# Audiovisual Speech Synthesis Based on Hidden Markov Models

Dipl.-Ing. Dietmar Schabus, Bakk.techn.


Dissertation
Submitted for the degree of Doctor of Technical Sciences

Graz University of Technology
Institute of Signal Processing and Speech Communication

Supervision: Assoc.-Prof. Dipl-Ing. Dr.mont. Franz Pernkopf
Co-supervision: Mag.phil. Dr.techn. Michael Pucher

Examiners:
Assoc.-Prof. Dipl-Ing. Dr.mont. Franz Pernkopf
Assoc.-Prof. B.Eng. M.Eng. Dr.Eng. Junichi Yamagishi

Graz, October 2014

## Declaration

I declare that I have authored this dissertation independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral dissertation.

## Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

Datum / Date

Unterschrift / Signature

# Abstract

In this dissertation, new methods for audiovisual speech synthesis using Hidden Markov Models (HMMs) are presented and their properties are investigated. The problem of audiovisual speech synthesis is to computationally generate both audible speech as well as a matching facial animation or video (a "visual speech signal") for any given input text. This results in "talking heads" that can read any text to a user, with applications ranging from virtual agents in human-computer interaction to characters in animated films and computer games.

For recording and playback of facial motion, an optical marker-based facial motion capturing hardware system and 3D animation software are employed, which represent the state of the art in the animation industry. For modeling the acoustic and motion parameters of the synchronously recorded speech data, an existing HMM-based acoustic speech synthesis framework has been extended to the visual and audiovisual domains.

The most important scientific contributions are on the one hand a novel joint audiovisual approach, where speech and facial motion are generated from a single model which combines both modalities. An analysis of the resulting HMMs and subjective perceptual experiments show that this way of modeling results in better synchronization between speech and motion than separate acoustic and visual modeling, which is the most commonly followed strategy in related work. On the other hand, average voice training and target speaker adaptation are investigated for the visual domain. The concept of adaptation has been one of the key factors for the popularity of the HMM-based framework for acoustic speech synthesis. Again, objective analysis and subjective perceptual experiments show that this concept is also applicable to the visual domain.

In order to study these modeling approaches, suitable data collections are required. To this end, several synchronous labeled corpora of speech and facial motion recordings in Austrian German have been created as part of this dissertation. The resulting data collections have been released on the Internet for research purposes, and may turn out to be valuable resources for the scientific community.

# Kurzfassung

In dieser Dissertation werden neue Methoden für audiovisuelle Sprachsynthese unter Verwendung von Hidden Markov Modellen (HMMs) präsentiert und auf ihre Eigenschaften untersucht. Die Problemstellung der audiovisuellen Sprachsynthese besteht darin, computerbasiert für eine beliebige textliche Eingabe sowohl hörbare Sprache als auch eine damit zusammenpassende Animation oder ein Video eines Gesichts (ein "visuelles Sprachsignal") zu erzeugen. Dabei entsteht ein "sprechender Kopf", der in der Lage ist, einem Benutzer beliebige Texte vorzulesen. Mögliche Anwendungen davon reichen von sogenannten virtuellen Agenten in der Mensch-Maschine-Interaktion bis zu computergesteuerten Figuren in Animationsfilmen und Computerspielen.

Zur Aufzeichnung und Wiedergabe von Gesichtsbewegungen wurden ein optisches, Marker-basiertes Aufzeichnungsgerät für Gesichtsbewegungen bzw. 3D Animations-Software verwendet, welche dem Stand der Technik in der Animations-Branche entsprechen. Für die Modellierung der akustischen sowie der Bewegungs-Parameter, die aus den synchron aufgezeichneten Sprachdaten gewonnen wurden, wurde ein bestehendes HMM-basiertes akustisches Sprachsynthese-System auf den visuellen und audiovisuellen Bereich erweitert.

Die wichtigsten wissenschaftlichen Errungenschaften sind zum einen ein neuer kombiniert-audiovisueller Ansatz, bei dem Sprache und Gesichtsbewegungen aus einem einzigen Modell generiert werden, das beide Modalitäten kombiniert. Eine Analyse der resultierenden HMMs und subjektiv-perzeptive Experimente zeigen, dass diese Art der Modellierung zu besserer Synchronisierung zwischen Sprache und Bewegung führt als separate akustische und visuelle Modellierung. Zum anderen werden Durchschnitts-Modell-Training und Zielsprecher-Adaption für den visuellen Bereich untersucht. Das Konzept der Adaption hat wesentlich zur Popularität des HMM-basierten Ansatzes für akustische Sprachsynthese beigetragen. Auch hier zeigen eine objektive Analyse und subjektiv-perzeptive Experimente, dass dieses Konzept auch auf den visuellen Bereich anwendbar ist.

Zur Untersuchung dieser Modellierungs-Ansätze werden entsprechende Da-

tensammlungen benötigt. Deswegen wurden im Zuge dieser Dissertation mehrere synchrone, annotierte Korpora bestehend aus Aufnahmen von Sprache und Gesichtsbewegungen in österreichischem Deutsch produziert. Die so entstandenen Datensammlungen wurden im Internet für Forschungszwecke veröffentlicht, und könnten sich als wertvolle Ressourcen für die wissenschaftliche Gemeinschaft herausstellen.

# Acknowledgments

First and foremost I want to thank Dr. Michael Pucher, who has been my mentor, FTW-internal PhD supervisor, and project manager since I joined FTW to work on my master's thesis in 2008. His great proficiency in winning research grants allowed me to work on a PhD in this interesting field in the first place, and his scientific guidance has been invaluable over the years.

I am very grateful to Prof. Franz Pernkopf, who supervised my PhD at TU Graz and did so in a very open and approachable manner, accepting the somewhat atypical setup of me working on a PhD at FTW and hence outside of the university, and who gave very valuable feedback at various stages of this undertaking.

I want to thank Dr. Gregor Hofer, for developing the idea (and the project proposal) for this research together with Dr. Pucher, for sharing his expertise on facial animation, for fruitful discussions and for helping me improve my papers.

Many thanks go to Prof. Junichi Yamagishi for accepting to act as second examiner for my dissertation and its defense.

I am thankful to Dr. Markus Kommenda, for his lecture on speech input and output at TU Vienna which sparked my interest in the field, and for opening the door to FTW (which he headed at the time) for me.

Furthermore, I am grateful to my long-time office mates and PhD companions Markus Toman, Andreas Sackl and Dr. Matthias Baldauf, as well as to many other colleagues at FTW, for contributing to a productive, yet friendly and open work atmosphere, with room for occasional fun.

On a less personal level, I want to acknowledge the funding of the research project "Adaptive Audio-Visual Dialect Speech Synthesis" by the Austrian Science Fund (FWF) under the project number P22890-N23.

Finally, on an entirely personal level, I am grateful to my love Marion, my friends and my family for their support and encouragement.

# Contents

# List of Acronyms and Abbreviations

**AMTV** Acoustic Modeling and Transformation of Varieties for Speech Synthesis 19

**AVDS** Adaptive Audio-Visual Dialect Speech Synthesis 19, 100

**CMLLR** Constrained Maximum Likelihood Linear Regression 74, 136

**CNRS** National Center for Scientific Research 40

**CSMAPLR** Constrained Structural Maximum A Posterioi Linear Regression 74

**CSTR** Centre for Speech Technology Research 97, 115, 116, 136

**EM** Expectation Maximization 56, 57

**EMA** Electro-Magnetic Articulography 99, 102–105, 108

**EMIME** Effective Multilingual Interaction in Mobile Environments 97, 115, 116, 136

**F0** Fundamental Frequency 60–62, 67, 115–117, 119, 123, 124, 132

**FMSC** Face Motion and Speech Corpus 100, 102, 103, 111, 137

**FTW** *Forschungszentrum Telekommunikation Wien* 19, 57

**GIDS** Bad Goisern and Innervillgraten Dialect Speech Corpus 100, 102, 103, 111, 115

**HMM** Hidden Markov Model 17, 19, 20, 22, 27–29, 33–35, 41–44, 47, 49, 52–57, 59–62, 66, 71, 84, 86, 114, 116, 135, 136, 143–145

**HSMM** Hidden Semi-Markov Model 56–58, 62, 63, 67–73, 95–98, 113, 115, 116, 118, 123, 136

**HTK** Hidden Markov Model Toolkit 27, 63, 105

| | |
|---|---|
| **HTS** | HMM-based Speech Synthesis System 27, 33, 41, 52, 56, 57, 60, 62, 64–68, 81, 97, 115, 116, 123, 136, 144 |
| **KTH** | Royal Institute of Technology 37–40 |
| **MDL** | Minimum Description Length 66 |
| **MFCC** | Mel-Frequency Cepstral Coefficient 34, 50–53, 67, 124, 132 |
| **MLLR** | Maximum Likelihood Linear Regression 72–74 |
| **MMASCS** | Multi-Modal Annotated Synchronous Corpus of Speech 102, 111 |
| **MSD** | Multi-Space Distribution 60–62, 115 |
| **PCA** | Principal Component Analysis 33, 34, 40, 41, 46, 84, 86, 87, 89, 90, 95, 96, 115–117, 123, 139, 144, 146 |
| **PDF** | Probability Density Function 54, 55, 57, 58, 61 |
| **PEC** | Phoneme Equivalence Class 116, 118, 119 |
| **RMSE** | Root Mean Squared Error 88–90, 93, 139, 140 |
| **SALB** | Speech Synthesis of Auditory Lecture Books for Blind Children 19, 57 |
| **SAT** | Speaker-Adaptive Training 72, 74, 136 |
| **SVD** | Singular Value Decomposition 86–90, 96, 137, 139 |
| **TTS** | Text-To-Speech 23, 24, 26, 34, 41, 46, 47, 64, 74, 80, 115 |
| **UCSC** | University of California at Santa Cruz 37–39 |
| **VSDS** | Viennese Sociolect and Dialect Synthesis 19 |

# Chapter 1

# Introduction

This dissertation addresses the problem of audiovisual speech synthesis, i.e., the problem of computationally generating both audible spoken language (speech), as well as corresponding facial movement ("visual speech") for a given textual input (as illustrated by Figure 1.1). Speech synthesis in the *acoustic* domain has been an active research area for more than 50 years, and synthetic voices have become commonplace in application fields like car navigation devices, smartphones, public transport announcements, automated telephone services, and assistive technologies. Adding a speaking face via *visual* speech synthesis can be appropriate in speech-based human-computer interfaces because of improved intelligibility and increased attention and engagement. Furthermore, flexible computer-controlled talking characters are useful in the entertainment industry for the creation of animated films and computer games. Visual speech synthesis is younger and less settled than its acoustic counterpart, but nonetheless a well established research field which has seen much progress since its beginnings.

Often, the techniques used for visual speech synthesis are the same or similar as in acoustic synthesis, and this is also true for the work presented here. In particular, an approach based on Hidden Markov Models (HMMs) has received much scientific attention since the early 2000s in acoustic speech synthesis, and it has already been applied in various ways for visual speech synthesis as well. In this approach, a speech model is "learned" from a collection of speech recordings during a "training" phase, and the resulting model can then be used at synthesis time to generate speech for any new input sentence. In this dissertation, the HMM-framework is applied for both acoustic and visual speech synthesis, for two reasons: first, this allows for true joint modeling of the two modalities in a single multi-modal model, and second, there are interesting "advanced" modeling techniques like speaker adaptation which have played an important role for the success of the HMM-

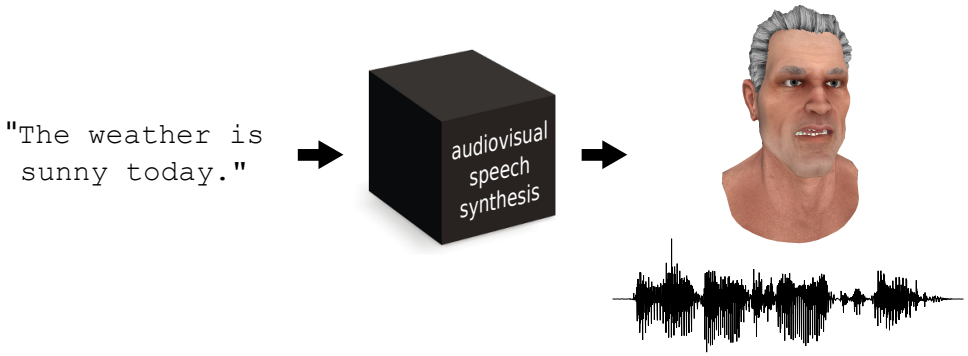



Figure 1.1: An audiovisual speech synthesis system takes text as input, and produces speech and matching facial movement as output.

framework in the acoustic domain, but which have not been investigated before for the visual domain.

For the research in this dissertation, the decision was made to represent the visual speech signal using a 3D head model, rather than using video recordings, for example. Visual speech synthesis is therefore realized as generating motion control parameters for a 3D head model. The control parameters thereby consist of the movement of a number of facial feature points. On the one hand, this allows the collection of recordings from human speakers by tracking markers that are glued to their faces. On the other hand, it defines a speaker-independent control parameter set, thus allowing experiments using data from multiple speakers, such as speaker adaptation.

Because speech synthesis is a subfield of signal processing, it has traditionally been assigned mostly to electrical engineering. With increasing computational power and the trend towards data-driven methods, other elements that are somewhat closer to computer science became important, like combinatorics (for handling the large number of possible combinations resulting from a large collection of data), statistical machine learning and computational linguistics. For adequate modeling of speech, an understanding of some concepts from phonetics and linguistics are necessary. Visual speech synthesis additionally brings computer graphics and/or image processing, and even computer vision (for recording speakers' facial movements) into the picture. Finally, human perceptional factors need to be considered, since the aim of this kind of technology is to produce speech that is convincing to human viewers/listeners. Audiovisual speech synthesis is therefore a highly interdisciplinary field, which this dissertation tries to account for by touching on all of the mentioned topics (to varying degree).

## 1.1 Project Background

The research in this dissertation was conducted as part of the project Adaptive Audio-Visual Dialect Speech Synthesis (AVDS), carried out at the *Forschungszentrum Telekommunikation Wien* (FTW) (Telecommunications Research Center Vienna) in Vienna, Austria. FTW is a non-profit research institution that focuses on applied research in cooperation with academic and industrial partners, but also carries out basic research projects (like AVDS), in the field of information and communication technology. The AVDS project was funded by the *Fonds zur Förderung der wissenschaftlichen Forschung* (FWF, Austrian Science Fund) under the project number P22890-N2. Project manager and principal investigator was Michael Pucher of FTW, the project lasted from January 2011 to December 2014. During this time, the author of this dissertation was employed as a full-time researcher at FTW, working on AVDS and several other FTW research projects. Within AVDS, FTW collaborated with Sylvia Moosmüller and others of the Acoustics Research Institute of the Austrian Academy of Sciences.

Speech technology has been a permanent research topic at FTW since the center's foundation in 1998, beginning with the collection of an Austrian German telephone speech corpus (M. Baum et al., 2000). The first work on speech synthesis at FTW was carried out by Michael Pucher and others (Pucher et al., 2003) in the "Speech & More" project (1999–2003), which was led by Georg Niklfeld. The first FTW research project concentrating on synthesis was Viennese Sociolect and Dialect Synthesis (VSDS) from 2007 to 2009, in which the author of this dissertation worked on his master's thesis (Schabus, 2009). VSDS was also led by Michael Pucher and can be seen as the direct predecessor project to AVDS. In parallel to AVDS, two further projects focusing on speech synthesis have been running at FTW, Acoustic Modeling and Transformation of Varieties for Speech Synthesis (AMTV) and Speech Synthesis of Auditory Lecture Books for Blind Children (SALB), to both of which the author of this dissertation has made some contributions.

## 1.2 Scientific Contributions

In this dissertation, new methods are developed for modeling visual and audiovisual speech using the HMM-based framework, making the following scientific contributions:

- Joint audiovisual modeling, using a single combined model for both modalities, is proposed as a simple but effective approach for ensuring synchronization between generated speech and facial motion. The dif-

ferences to several strategies of separate acoustic and visual modeling are analyzed objectively and subjective evaluations show a noticeable improvement in synchrony. Importantly, the quality of the generated acoustic speech is not negatively affected by combined modeling.

- Average voice training and target speaker adaptation, an important advantage of the HMM-based speech synthesis framework, have been applied to visual speech parameters, showing that this concept is also applicable to this domain. As for acoustic speech synthesis, the adaptive approach is able to outperform the single-speaker approach, when the amount of training data from the target speaker is small.
- A feature extraction method for the visual data based on principal component analysis is proposed and shown to result in features that are not only well suited for statistical modeling, but also for finding a reduced space that is still general enough to "contain" new target speakers, making the features suitable also for speaker-adaptive modeling.
- A pipeline from recording via feature extraction and model training to synthesis and final animation has been developed and is described in detail. For the modeling part, an existing HMM-based framework for acoustic speech synthesis was extended to include the visual modality.
- Several synchronous multi-modal speech corpora have been created, using marker-based optical facial motion tracking for the visual modality, studio-quality audio for the acoustic modality, and (for part of the data) electromagnetic articulography for tongue motion tracking. Data from eleven speakers in Standard Austrian German, from eight speakers in two different Austrian dialects, and from one speaker speaking at normal, slow and fast speaking rate has been recorded, preprocessed and manually refined. All resulting data has already been or will be released to the research community on the Internet.

## 1.3 Publications

Parts of this dissertation have been previously published in the journal articles and conference papers listed below, each of which needed to pass a peer-reviewing process with at least two reviewers assessing the quality of the submitted manuscript. All conference papers were also presented as oral or poster presentations at the respective conference by the author, except for paper number 3, which was presented by Michael Pucher. The relation between these publications and this dissertation will be indicated at the respective places.

### Journal Articles

1. D. Schabus, M. Pucher, and G. Hofer (Apr. 2014a). "Joint Audiovisual Hidden Semi-Markov Model-based Speech Synthesis". In: *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347. URL: http://dx.doi.org/10.1109/JSTSP.2013.2281036

### Conference Papers

2. D. Schabus, M. Pucher, and G. Hofer (Sept. 2012b). "Speaker-adaptive visual speech synthesis in the HMM-framework". In: *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, pp. 979–982. URL: http://www.isca-speech.org/archive/interspeech_2012/i12_0979.html

3. D. Schabus, M. Pucher, and G. Hofer (May 2012a). "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 3313–3316. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/302.html

4. D. Schabus, M. Pucher, and P. Hoole (May 2014b). "The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland, pp. 3411–3416. URL: http://www.lrec-conf.org/proceedings/lrec2014/summaries/192.html

5. D. Schabus, M. Pucher, and G. Hofer (Sept. 2013). "Objective and Subjective Feature Evaluation for Speaker-Adaptive Visual Speech Synthesis". In: *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing (AVSP)*. Annecy, France, pp. 37–42. URL: http://avsp2013.loria.fr/proceedings/papers/paper_31.pdf

6. D. Schabus, M. Pucher, and G. Hofer (Aug. 2011). "Simultaneous speech and animation synthesis". In: *ACM SIGGRAPH Posters, 38th International Conference and Exhibition on Computer Graphics and*

*Interactive Techniques*. Vancouver, BC, Canada, 8:1–8:1. URL: http://dx.doi.org/10.1145/2037715.2037724

7. M. Pucher, D. Schabus, G. Hofer, N. Kerschhofer-Puhalo, and S. Moosmüller (Feb. 2012). "Regionalizing Virtual Avatars – Towards Adaptive Audio-Visual Dialect Speech Synthesis". In: *Proceedings of the 5th International Conference on Cognitive Systems (CogSys)*. Vienna, Austria, pp. 1–1

8. J. Hollenstein, M. Pucher, and D. Schabus (Sept. 2013). "Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis". In: *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing (AVSP)*. Annecy, France, pp. 31–36. URL: http://avsp2013.loria.fr/proceedings/papers/paper_20.pdf

## 1.4 Overview of This Dissertation

The remainder of this dissertation is organized as follows. Chapter 2, titled Audiovisual Speech Synthesis Background, discusses the problems of acoustic and visual/audiovisual speech synthesis and the most important approaches to solving them. Especially for the visual/audiovisual domain it attempts to provide a broad overview over related work. Chapter 3, titled Speech Synthesis Using Hidden Markov Models, is intended to give an understanding of how speech synthesis using HMMs works, focusing on the audio-only case. Chapter 4, titled Developing an Audiovisual Speech Synthesis Pipeline, gives all details about the visual recordings, visual features and how an existing acoustic HMM-based speech synthesis system was extended to visual and audiovisual modeling. Chapter 5, titled Audiovisual Speech Corpora, describes the speech data collections created. Chapter 6, titled Synchronization of Speech and Motion, describes and compares different methods of synchronizing speech and facial motion generated from HMMs, and argues for a combined single model for both modalities. In Chapter 7, titled Speaker-Adaptive Audiovisual Speech Synthesis, the concept of average voice training and target speaker adaptation is applied to the visual domain. Finally, Chapter 8, titled Conclusion, summarizes the most important findings and gives an outlook to possible future research related to this body of work.

# Chapter 2

# Audiovisual Speech Synthesis Background

The aims of this chapter are to define central terminology and concepts, to motivate the research presented in the following chapters and to provide an overview over related work. First, speech synthesis in general terms and concerning the acoustic modality is discussed. Then, the problem of speech synthesis is extended to the visual domain, and several approaches to this multi-modal problem are presented.

## 2.1 Speech Synthesis in the Acoustic Domain

In the author's master's thesis (Schabus, 2009), speech synthesis was introduced using the following two paragraphs:

> The main means of human communication is speech. To enhance human-machine interaction, computers should be capable of speech communication. Among other things, this requires computers to be able to produce an acoustic speech signal for a given input text, i.e., pass information to a human user through "speaking". The scientific field of speech synthesis deals with the development of Text-To-Speech (TTS) systems that satisfy this requirement.
>
> Besides the attempt to make human-machine communication more natural, speech synthesis has applications wherever visual output has disadvantages. For example, acoustic output is suitable for a user steering a vehicle or aircraft (or other activities involving eyes and hands), it can be used over the well-established

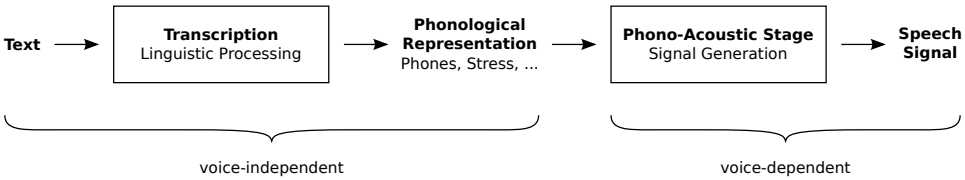



Figure 2.1: Separation of a speech synthesis system into a voice-independent and a voice-dependent part (figure after Pfister and Kaufmann, 2008).

> telephone network (for information, reservation or ordering services), and it easily attracts our attention (for public announcements, alarm systems, etc.). Furthermore, speech synthesis has long been a vital assistive technology for people with visual impairment or other reading difficulties (e.g., screen readers), as well as for people with speech impairment (e.g., voice output communication aids).

More elaborate and complete introductions to speech synthesis can be found in the books of Dutoit (1997), Furui (2001), or Pfister and Kaufmann (2008), for example, but the above two short paragraphs already cover several important aspects. One key point is that the input to a TTS system is text and hence discrete and symbolic, whereas the output is a continuous speech signal which has many concrete properties that are not part of the input, like voice characteristics. Pfister and Kaufmann (2008) therefore divide TTS systems into two parts along this border of text-related versus signal-related, as illustrated in Figure 2.1.

The first part (transcription) is responsible for translating the input text into a phonological representation that is still purely symbolic, but gives specific and complete information on how to pronounce the input sentence(s). To do so, this part usually employs a pronunciation dictionary and a collection of letter-to-sound rules to determine the sequence of speech sounds (phones) and which syllables to stress. Additionally, it may produce information on intonation, sentence stress, and phrasing, based on a syntactic analysis of the sentence, on the type of utterance (statement, question, exclamation), and other things that can be derived from text.

The second part (called the phono-acoustic stage by Pfister and Kaufmann, 2008) takes the phonological representation produced by the transcription part as input and generates an appropriate speech signal for it. There are several possible ways to achieve this, some of which will be discussed shortly, but they all have in common that they produce speech in a certain voice, with all its specifics, such as speaker identity, gender, approximate age, loudness, speaking rate, regional and/or social language variety, etc. The first part on the other hand is independent of the voice used later, in fact it

is even independent of the way the phono-acoustic stage is realized.

Several approaches have been developed for the phono-acoustic stage in the second half of the 20th century. The earlier among them are covered, e.g., by Dutoit (1997), Furui (2001), and Pfister and Kaufmann (2008). Here, only the two methods prevalent today shall be discussed: Speech synthesis based on unit selection and concatenation, and speech synthesis based on parameter statistics, the latter being the method applied in the research for this dissertation. Both of them, as well as hybrid systems which combine the two methods, have been under active research from the 1990s until today.

### 2.1.1 Speech Synthesis via Waveform Concatenation

The basic idea of unit selection systems (Sagisaka et al., 1992; Hunt and Black, 1996) is to concatenate speech signal cut-outs (units), which are taken from a collection of recorded utterances, thus forming new combinations, to synthesize new utterances. The assumption is that appropriate units for any utterance are available, which is true if the amount of collected recordings is large enough. The units in the system described by Hunt and Black (1996) are phones, but differently sized units are also possible. To synthesize a new utterance, appropriate units are selected from the collection based on the minimization of a cost function, using the Viterbi Algorithm (Viterbi, 1967). After transcription of the input text, the phonological representation consists of a sequence of symbolic target units $t_{1:n} = (t_1, ..., t_n)$, for each of which a concrete speech unit $u_i$ from the recording collection needs to be selected. These selected units $u_{1:n} = (u_1, ..., u_n)$ are then concatenated to form the result. The cost function consists of a *target cost* term, expressing how well a candidate unit from the collection satisfies the requirements of the target unit (by comparing speech units to a symbolic unit), as well as a *concatenation cost* term, which expresses how well two consecutive candidate units combine (by comparing two speech units).

The target cost between a candidate unit $u_i$ and the required target unit $t_i$ is defined as

$$C^{(t)}(t_i, u_i) = \sum_{j=1}^{p} w_j^{(t)} \, C_j^{(t)}(t_i, u_i), \tag{2.1}$$

which is a weighted sum of the differences of the candidate unit and the target unit regarding $p$ phonetic and prosodic features. Such sub-costs express phonetic and prosodic properties of the unit in question, as well as of its context (the preceding and succeeding units, e.g.). Hunt and Black (1996) report to have used 20–30 such features.

The concatenation cost between two successive candidate units $u_{i-1}$ and $u_i$ is defined as

$$C^{(c)}(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^{(c)} C_j^{(c)}(u_{i-1}, u_i), \tag{2.2}$$

which is also a weighted sum of $q$ sub-costs, which in this case express discontinuities at the point of concatenation. Hunt and Black (1996) used cepstral distance, difference in power and difference in pitch, i.e., $q = 3$. For a candidate sequence $u_{1:n}$, the total cost is then the sum of all $n$ target costs plus the sum of all $n - 1$ concatenation costs:

$$C(t_{1:n}, u_{1:n}) = \sum_{i=1}^{n} C^{(t)}(t_i, u_i) + \sum_{i=2}^{n} C^{(c)}(u_{i-1}, u_i). \tag{2.3}$$

For the final result, the unit sequence $\hat{u}_{1:n}$ that minimizes this total cost is found via dynamic programming, and the units are concatenated to produce the output signal, i.e.,

$$\hat{u}_{1:n} = \operatorname*{argmin}_{u_{1:n}} C(t_{1:n}, u_{1:n}). \tag{2.4}$$

Different unit selection systems differ mainly in what features are used in the cost functions, and how the weights for them are chosen or computed. By their nature, concatenative systems strongly rely on the quality of the speech data collection. Furthermore, the amount of data for a high-quality voice needs to be quite large, and this large amount of data is required at run-time. The very high quality of synthetic speech that can be achieved with the unit selection technique has made it common in commercial TTS systems (Breen and Jackson, 1998; Donovan and Eide, 1998; Beutnagel et al., 1999; Coorman et al., 2000).

### 2.1.2 Statistical Parametric Speech Synthesis

In statistical parametric speech synthesis (Zen et al., 2009; Tokuda et al., 2013), the produced speech signal does not (directly) contain any recorded speech. Instead, this approach to the phono-acoustic stage task relies on an analysis–re-synthesis technique. From a recorded speech utterance, a sequence of speech parameters can be extracted using the analysis procedure, and the re-synthesis procedure can be used to turn this parameter sequence back into a speech waveform again. Using the extracted parameter sequences of a collection of recorded utterances as training data, statistical models for these speech parameters are trained, typically using a maximum likelihood criterion to optimize the model parameters for the given data. To synthesize a new utterance, the most probable speech parameter sequence for the text

input is generated from the statistical models, and finally the re-synthesis procedure produces the result waveform from this sequence.

Although any generative model could be used in this setup, HMMs are typically used, also in this dissertation. Chapter 3 discusses HMM-based speech synthesis in detail, in the following only a few relevant points shall be made.

HMMs became widespread in speech recognition in the 1970s and 1980s (Baker, 1975; Jelinek et al., 1975; Poritz, 1982; Juang and Rabiner, 1985), and rapidly gained popularity in speech synthesis in the 1990s, largely due to a series of publications from the Nagoya Institute of Technology and the Tokyo Institute of Technology (Tokuda et al., 1995; Masuko et al., 1996; Yoshimura et al., 1998; Tokuda et al., 1999; Yoshimura et al., 1999), as well as to the release of the HMM-based Speech Synthesis System (HTS)[1] in 2002 by the same people (Zen et al., 2007a). The HTS system is actually an extension of another system, the Hidden Markov Model Toolkit (HTK)[2] (Woodland et al., 1994; Young et al., 2006), which was primarily designed for speech recognition.

Similar to most speech recognition and also unit selection systems, HMM-based speech synthesis systems typically use phones as modeling units, i.e., an utterance like "There was a change now" is essentially represented as a sequence of phones, e.g., $(dh, eh, r, w, aa, z, ax, ch, ey, n, jh, n, aw)$. Each element of the sequence usually holds a phone symbol (e.g., $ey$ to represent the diphthong in the word "change"), its phonetic context (e.g., the $ey$ is preceded by an $ax$ and a $ch$, and succeeded by an $n$ and a $jh$), and additional information like whether or not the corresponding syllable is stressed, etc. As this representation is intended as an annotation of recorded (time-variant) speech signals, the begin and end times are also normally given for each element. These phone borders are either manually labeled, or determined automatically via forced alignment (as described in Section 3.2) or other automatic alignment methods, possibly followed by manual corrections. From all the phone speech signal segments defined by these borders on the training data (collection of recorded utterances), a collection of phone HMMs is trained, where—in a simplified view—each HMM captures the average sound of all signal segments from the training data that were used to train this specific HMM.

As already mentioned, these phone HMMs are not trained on the raw speech waveform itself (a one-dimensional signal sampled at, e.g., 44 100 Hz), but on a parametric representation of it, which results from the analysis procedure of the analysis–re-synthesis technique. This representation consists of de-

[1] http://hts.sp.nitech.ac.jp
[2] http://htk.eng.cam.ac.uk

correlated higher-dimensional features at a lower sampling rate; for example, the 150 ms *ey* segment is represented by 23 41-dimensional vectors rather than by 6615 scalar values, i.e., the sampling rate changes from 44 100 Hz to 200 Hz. These feature vectors then serve as training data for the phone HMMs.

In order to synthesize an utterance given as text, the phonological representation (in particular, the phone sequence) is determined by a transcription module, such that the appropriate phone HMMs can be found and concatenated to an utterance HMM. From this, the most probable sequence of feature vectors is computed via a trajectory generation algorithm (Tokuda et al., 1995). Finally, this sequence is turned into a speech signal using the re-synthesis procedure.

Compared to unit selection methods, HMM-based speech synthesis has several advantages (Zen et al., 2009; Tokuda et al., 2013), mainly due to the fact that unit selection systems are inherently limited to produce speech that sounds like the recorded material. In an HMM-based system, on the other hand, voice characteristics, speaking styles, emotions, etc. can be changed by modifying the parameters, for example by using model adaptation (discussed in Section 3.5) or interpolation (e.g., Yoshimura et al., 2000). The parametric approach thus offers more flexibility. Furthermore, unit selection requires large amounts of speech recordings for high quality voices. This is generally also true for the HMM-based approach, however, using the adaptation technique it is possible to train an "average voice" on a large amount of data from multiple speakers and then adapt the models towards a target speaker with only a small amount of data from that speaker. Like HMM-based synthesis in general, also the idea of speaker-adaptation rooted in the speech recognition field. Finally, HMM voices require significantly less disk storage and memory space than unit selection voices, because for synthesis only the statistical models are needed, whereas a unit selection system requires the entire collection of recordings at run-time.

The main disadvantage of HMM-based speech synthesis in comparison to unit selection is the lower achieved quality of synthesized speech. Zen et al. (2009) list three quality degrading factors: First, already analysis and re-synthesis of recorded speech exhibits a certain "buzziness" of the speech signal, due to the simple excitation model used (basically, a pulse train for voiced parts, and white noise for unvoiced parts). Second, in the statistical modeling of the parameters, several simplifying assumptions are made which do not really hold for real speech, e.g., piece-wise constant statistics within an HMM state. And third, the statistical averaging over observed instances in the training data to estimate the HMM parameters creates the problem of over-smoothing, resulting in a loss of detailed characteristics and a certain "muffledness" of the generated speech.

Nevertheless, HMM-based speech synthesis system have reached intelligibility scores similar to natural speech in the Blizzard Challenges (e.g., Karaiskos et al., 2008), a competition held every year since 2005 where different groups compare their speech synthesis systems built on the same data via subjective listening tests.

### 2.1.3 Hybrid HMM/Unit Selection Methods

Several systems have been developed which combine HMM-based speech synthesis and unit selection into "hybrid" systems, aiming for new concatenative synthesizers that use the statistical models to guide the selection of units. Some systems use the speech parameters predicted by HMMs as targets in the unit selection algorithm, thereby simplifying the target cost function (Kawai et al., 2004; Hirai and Tenpaku, 2004; Rouibia and Rosec, 2005; Yang et al., 2006), others use the likelihoods of candidate units, which are determined using HMMs, in the (target and/or concatenation) cost functions (X. Huang et al., 1996; Hon et al., 1998; Okubo et al., 2006; Ling and R.-H. Wang, 2007; Ling et al., 2007). Hybrid systems are able to reach a very high quality in synthetic speech (e.g., in the Blizzard Challenge 2012, King and Karaiskos, 2012), but it should be noted that not all advantages of the statistical parametric method can be retained, e.g., the flexibility in changing voices and the small footprint.

## 2.2 Speech Synthesis in the Visual Domain

After the overview over modern techniques for acoustic speech synthesis, we now turn to the visual domain. From a communications point of view, human speech in the acoustic domain is a "message" originating in the speaker's brain, "encoded" as a time-variant signal produced by the human vocal apparatus, transmitted as sound waves via the air, captured by the human ear and processed by the listener's brain, as described, e.g., by Coker et al. (1963) and illustrated in Figure 2.2. Switching to the visual domain, we might analogously look at human speech as an (additional) signal produced by the movement over time of the (visible parts of the) vocal apparatus, transmitted via light rays that originate from some light source, which are reflected from the surface of the speaker's face and which are finally captured by the viewer's eyes. Leaving aside facial expressions, eye and eyebrow movement, and other truly complementary "channels", one might argue that the visual signal of the articulator movement is merely a "byproduct" of the acoustic shaping of the speech sounds. For example, the rounding of the lips during production of the word "do" is purely acoustically motivated. However, it is intuitively clear that seeing the articulator movement of a

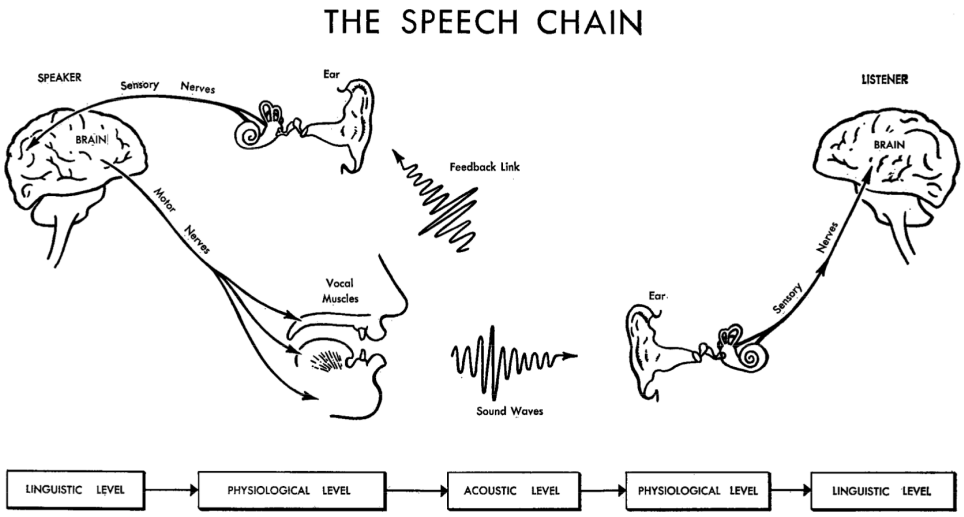



Figure 2.2: The "Speech Chain" (image from Coker et al., 1963).

speaker can help to disambiguate speech and thus improve intelligibility (especially under difficult acoustic conditions, see, e.g., Sumby and Pollack, 1954). Also, the presence of a face representing a speaking person (e.g., in an animated film or computer game) with speech accompanied by inappropriate articulator movement creates an undesirable inconsistency for the viewer. Depending on the situation, this "byproduct" can be of considerable importance. In any case, it is beyond question that the acoustic speech signal and the visible articulator movement originate from the same process and are very closely related.

On the technological side, we have microphones and loudspeakers for capturing and producing sound waves, and cameras and displays for capturing and reproducing visible light. However, a different, more animation-inspired viewpoint on visible speech motion is also possible: one that regards the facial motion itself as the relevant signal, not the light reflected from the surface of the moving face. If we think about a speech-capable humanoid robot, for example, which we want to be able to produce an appropriate visual speech signal when it speaks to a human user, then the robot's control unit would not need to take into account the lighting conditions or the human's viewpoint; it would not have to decide which color and brightness values to generate for an appropriate visual speech signal; it would simply need to move the robot's face "correctly", and all other things mentioned would follow from the laws of physics. This viewpoint on the visual speech signal is also valid in virtual three-dimensional worlds as commonplace in animated films, computer games, and the like, where speaking characters appear embedded in specific surroundings, with specific (virtual) lighting and seen from a specific viewpoint within the virtual space. In order to

have the flexibility to freely change such conditions as surroundings, lighting, viewpoint, etc., we need the visual speech generation algorithms either to take all of these conditions into account (simultaneously!), or we need the algorithms to be independent of these conditions. The latter is much easier to achieve, namely by creating visual speech in terms of facial motion, i.e., motion and deformation of the 3D character over time, and let the "physics" of the 3D scene rendering take care of the rest.

These two views on the visual speech signal (filming the visual speech signal vs. modeling and recreating it via rendering) give rise to two quite different approaches to the problem of synthesizing new visual speech signals. The first one operates on conventional video data, which can also be seen as image sequences, or as pixel intensity values changing over time. For recording the training data (in the case of a data-driven method), a conventional video camera can be used, and the system's output will also be a video. For the other approach, a way of recording speech motion in 3D and a way of "applying" such data to a 3D head model are required. Several possibilities have been proposed, but the end result will always be a motion and deformation sequence of the 3D head model, which can then be rendered as a video captured by a virtual camera, if necessary. The following two subsections present some relevant work for each of the two approaches. More elaborate introductions to (audio-)visual speech synthesis and surveys of developed methods can be found in the overview paper of Bailly et al. (2003), the PhD dissertation of Beskow (2003), and in the books of Parke and Waters (1996), Deng and Neumann (2008) and Bailly et al. (2012).

### 2.2.1 Visual Speech Based on Image Sequences

The general idea of most image-based visual speech synthesis systems is to assemble video frames taken from a collection of recorded footage of a speaking person to create new videos of that person speaking arbitrary new utterances. Often, this includes a position normalization step and a segmentation of the face into regions which are then treated separately. The other parts of the frame can then be treated as background video, for which no modification is required, and onto which the synthesized sequences of the foreground regions can be pasted.

In their "Video-Rewrite" system, Bregler et al. (1997) replace just the mouth region with a new image sequence selected from 8 minutes of recordings based on so-called tri-phones: for each phone in the new utterance, an image sequence is taken from the recordings that contains the same phone and also has the same preceding and succeeding phone, i.e., by considering a context of three phones' length. Figure 2.3 illustrates their face decomposition.[3] A

[3]Demonstration videos for "Video-Rewrite" are available at http://mrl.nyu.edu/

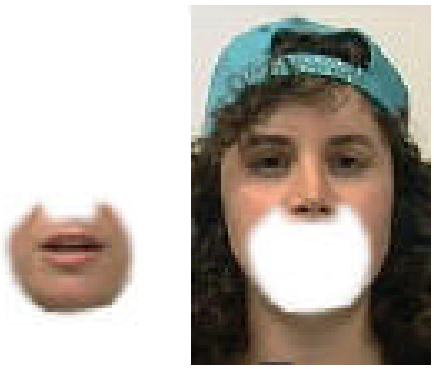


Figure 2.3: Separation into mouth region and background in "Video Rewrite", where the border of the region warps according to the automatically detected mouth and chin motions (image from Bregler et al., 1997).

Figure 2.4: Segmentation of the face into several parts for a concatenative image-based system (image from Cosatto, 2002).

more complex segmentation of the face was applied by Cosatto and Graf (2000) and Cosatto (2002), as shown in Figure 2.4. Their system was also developed further by Cosatto et al. (2000) and F. J. Huang et al. (2002) to apply the unit selection technique (cf. Section 2.1.1).

Instead of relying on the recorded data to contain every required frame, the system of Ezzat et al. (2002) also creates new frames using image morphing. Their system analyzes the available video data to automatically determine a set of key frames and morphing parameters, using the principal component analysis (Pearson, 1901; Shlens, 2014) and optical flow (Horn and Schunck, 1981) techniques. Figure 2.5 illustrates their system.[4]

~bregler/videorewrite/.

[4]Example videos for "Mary 101" are available at http://people.csail.mit.edu/

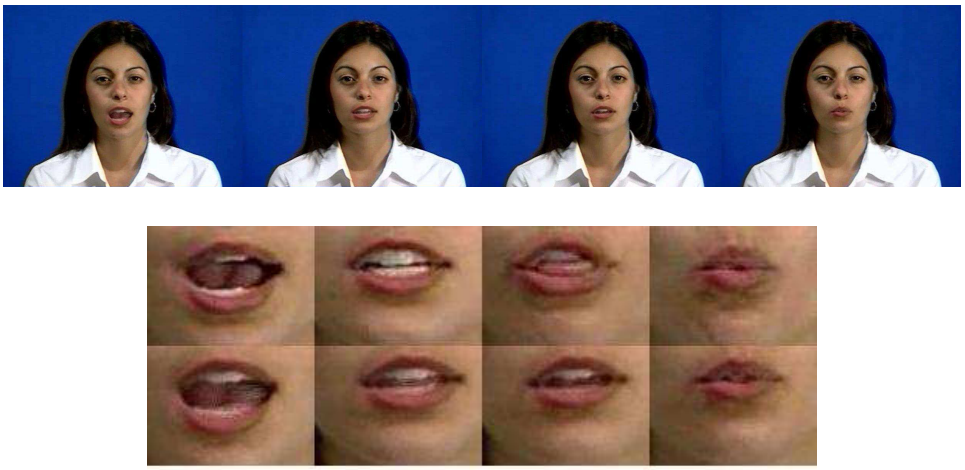


Figure 2.5: The image-based system "Mary 101". Top: Example output frames of the system, where the mouth region contains generated images, which are pasted into a background video with natural head and eye motion. Bottom: Comparison of original images from the corpus (top row) to corresponding system-generated images through morphing (bottom row). Images from Ezzat et al. (2002).

Theobald et al. (2002) proposed to train a shape and appearance model on video data for visual speech synthesis, where shape is defined as the position of tracked facial feature points and appearance as the images of the training data videos after shape normalization. They subsequently extended their system to 2.5D, meaning that the generated image sequence is applied as a dynamic texture map to a 3D face model, which deforms in two dimensions but not in the third, because the shape model computed from image data contains no depth information (Theobald et al., 2004). Figure 2.6 shows example frames from their system.[5]

The HTS working group has proposed an image-based visual speech synthesis system based on their HMM speech synthesis system (Sako et al., 2000). After applying Principal Component Analysis (PCA) to the pixel data of image sequences showing just the mouth region, phone HMMs were trained using the PCA-projected data as observation feature vectors. Acoustic speech is generated from a separate set of models; to synchronize the two outputs, the phone boundaries generated by the acoustic model are used for lip image synthesis. Figure 2.7 shows an example frame.[6]

---

tonebone/research/mary101/results/results.html.

[5] Demonstration videos available at http://www2.cmp.uea.ac.uk/~bjt/research/talking/demos.html.

[6] Demonstration video at http://www.mmsp.nitech.ac.jp/~sako/avi/sample1.avi.

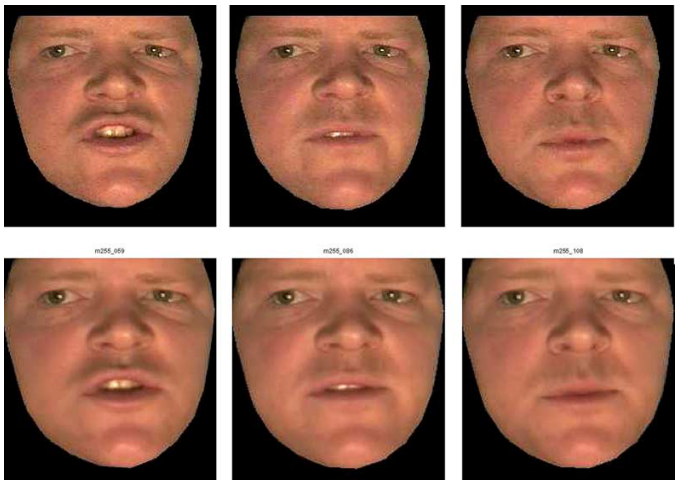


Figure 2.6: Illustration of the image-based system by Theobald et al., 2004. The top row shows images extracted from a recorded video sequence not used in training; the bottom row shows the corresponding output generated by the system (images from Theobald et al., 2004).

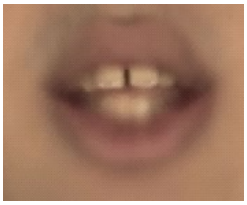


Figure 2.7: A mouth image resulting from PCA coefficients that were predicted from HMMs (video frame extracted from supplementary material published with Sako et al., 2000).

L. Wang et al. (2010) proposed an image-based visual speech synthesis system using a hybrid HMM/unit selection approach (cf. Section 2.1.3): First, audiovisual HMMs are trained to map acoustic to visual features.[7] Then, to generate new lip motion videos for a given audio speech sample (which may have been generated by a separate TTS module), a corresponding sequence of visual features is predicted from the HMMs, and finally an image sequence from the original database is selected based on these predictions, as illustrated in Figure 2.8. They refined their system to make use of the quite recent minimum generation error training procedure (L. Wang et al., 2011b), and extended their system to 2.5D by applying a dynamic texture to a static 3D head model (L. Wang et al., 2011a), as illustrated in Figure 2.9.[8]

[7] Concretely, Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) to image PCA features.

[8] Demonstration videos available at `http://research.microsoft.com/en-us/people/`

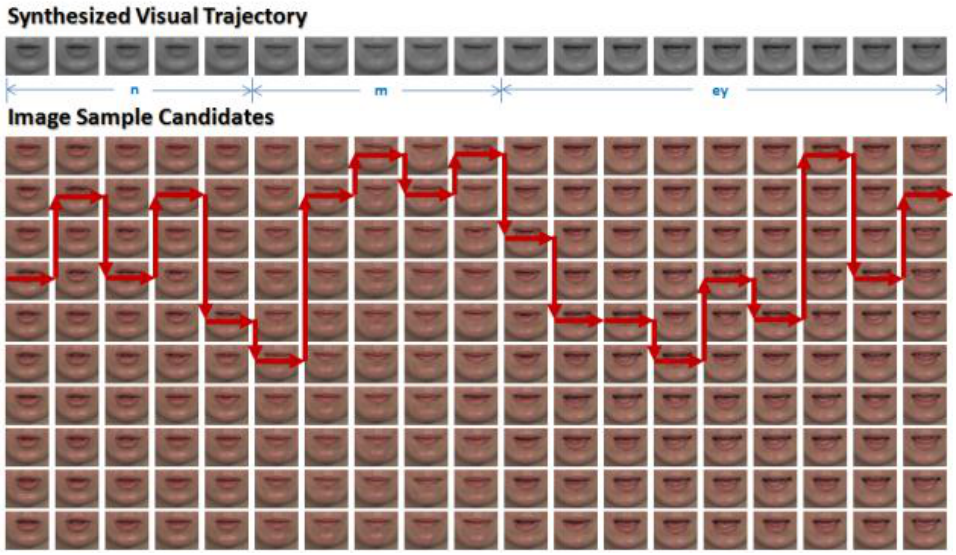



Figure 2.8: A sequence of mouth images that was predicted from HMMs (top) is used to guide the search for natural mouth images in the recorded collection (bottom). Image from L. Wang et al. (2010).

### 2.2.2 Visual Speech Based on 3D Head Model Deformation

The image-based systems presented in the previous subsection are capable of delivering impressive results, and they are ready to be applied in certain application fields, like synthesized newscasters, announcers, language tutors, virtual agents etc. However, for other applications such as animation films and computer games, they lack flexibility: The appearance of the speaker is given and fixed, but often it is desirable to have a different speaker (or an animal, a fantasy character, even an object) in the result. The viewing angle, field of view, lighting and background are also given and fixed, but it might be required to dynamically modify all of these. Even if some of the presented methods try to overcome the viewing angle limitation by applying a dynamic texture to a 3D face surface, this does not completely alleviate the shortcomings. The shading of the face cannot honor the specific lighting conditions defined by the concrete surroundings of a given 3D scene while the face deforms during articulation. Furthermore, the integration of a photo-realistic face into an otherwise 3D-rendered world of an animated film or computer game can be problematic.

These problems can be avoided by choosing a model-based approach to visual speech synthesis. Typically, the surface of the face is represented by

`lijuanw/`.

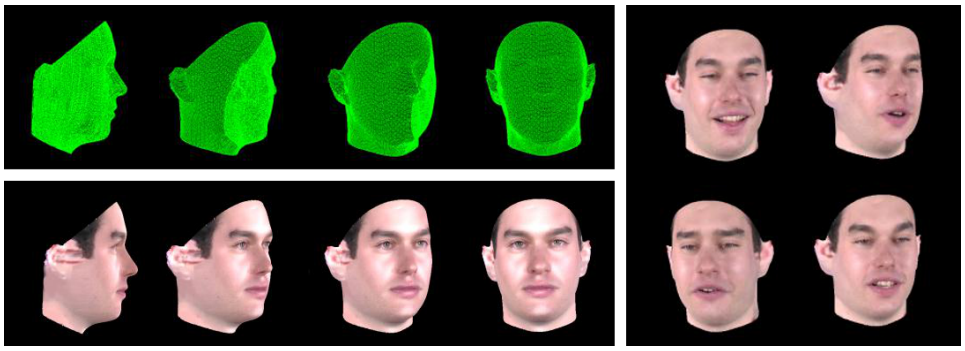


Figure 2.9: Simple 3D face model as wire-frame (top left) and with texture image applied (bottom left). Example frames with varying texture but static underlying 3D model (right). Images from L. Wang et al. (2011a).

a mesh of polygons in 3D space (in the same way as any other objects in the respective scene) and speech articulation (and other animation) is realized by moving the vertices of the mesh over time. The resulting dynamic 3D scene is then rendered repeatedly at equidistant points in time, resulting in an image sequence (video). During rendering, virtual lights placed in the scene, the position (and motion) of the virtual camera, surface materials and textures, and the respective models associated with these concepts (the rendering "physics") play an important role. Nevertheless, these topics from the field of computer graphics shall not be discussed here; for our purposes it is sufficient to assume such a computer graphics pipeline to be in place and to instead focus on the deformation of the 3D models alone.

It is worth noting that the model-based approach has been developed significantly earlier than the image-based approach. A compact control model for a not too fine-grained 3D mesh can result in a very efficient procedure for facial animation, tractable also on computers with much less computational power than what is commonplace today.

Groundbreaking work in this area dates back to the 1970ies. Frederick Parke created polygonal head representations by painting the polygon topology onto a person's face, and then reconstructing the 3D coordinates of the vertices by measuring their distances in multiple photographs, a method called photogrammetry, which is illustrated in Figure 2.10. These head models were then animated using key shape interpolation (Parke, 1972a; Parke, 1972b) and later using a parameter model combining interpolation, translation, rotation and scaling of various facial features (Parke, 1974; Parke, 1982). The parameters were chosen and refined manually, by studying video recordings and estimating the parameters to match the motion induced on the 3D head to the articulation seen in the video. Figure 2.11 shows example

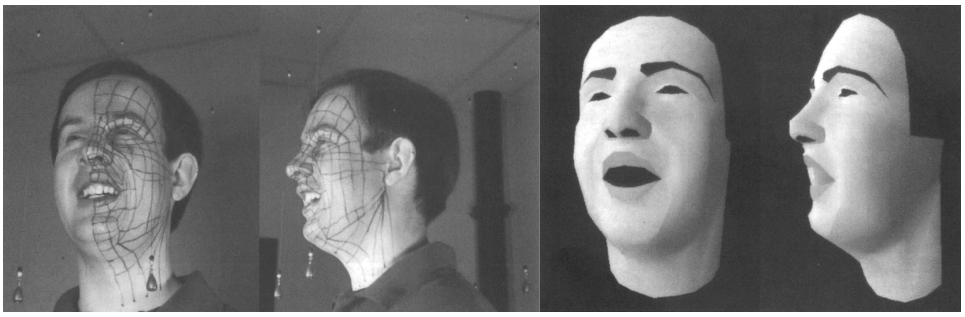


Figure 2.10: Reconstruction of a face in 3D by photogrammetry. The lines painted on the face facilitate establishing correspondences between multiple images, and also define the polygon topology of the 3D model (images from Parke, 1974).

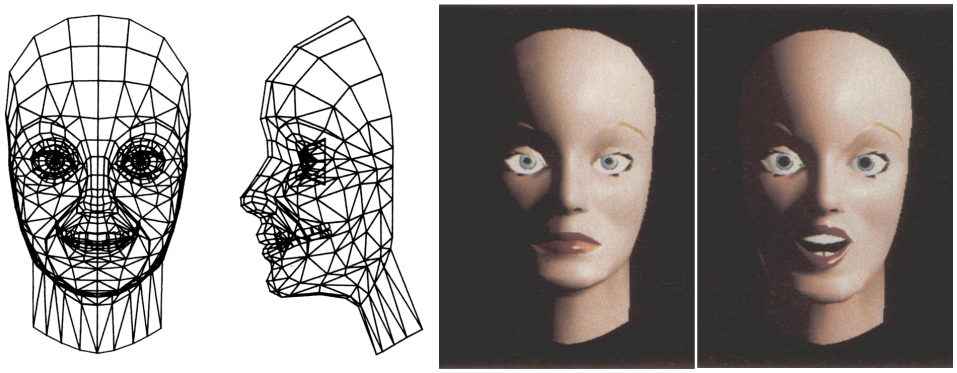


Figure 2.11: An early head model shown as wire-frame drawing (left) and as renderings with texture and shading, showing different facial expressions (right). Images from Parke (1982).

frames.[9]

Several research groups have created talking heads directly building on Parke's work, among them the Perceptual Science Lab of the University of California at Santa Cruz (UCSC) (Cohen and Massaro, 1993; Massaro, 1998; Cohen et al., 1998) and the Department of Speech, Music and Hearing of the Royal Institute of Technology (KTH) in Stockholm (Beskow, 1995; Beskow, 2004; Beskow, 2003), both of which have made many contributions to the field.

Parke's control model was modified by Pearce et al. (1986) to include additional speech-related control parameters, as well as to operate on input phoneme sequences. The resulting system was then further extended by

[9] An example video can be seen at http://www.youtube.com/watch?v=SPMFhcC4SvQ.

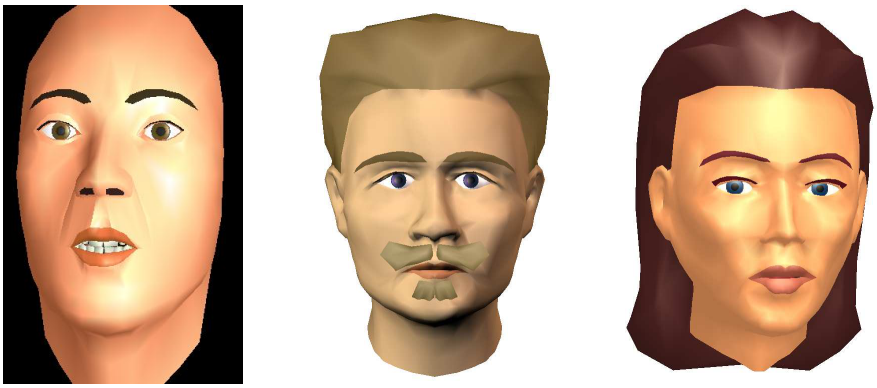


Figure 2.12: Talking head models based directly on the model from Parke (1974). From left to right: "Baldi" from UCSC and "August" and "Kattis" from KTH (images from Cohen et al., 1998 and Beskow, 2003).

Cohen and Massaro (1993) to include a simple tongue representation and to use dominance functions for addressing the problem of coarticulation, i.e., visible influences of neighboring phones on each other. A realistic palate, teeth and an improved tongue model based on 3D ultrasound data were subsequently added to this model (Cohen et al., 1998). Figure 2.12 (left) shows a frame from this system's output.[10] In order to directly map acoustic features derived from speech recordings to facial control parameters, Massaro et al. (1999) trained an artificial neural network on a corpus of parallel acoustic and visual parameter data. This data consisted of audio recordings of 400 isolated words for the acoustic part, and parameter sequences generated by their rule-based system, given the phoneme sequence and temporal borders of the audio recordings, for the visual part. Note that in this setup, no visual recordings were used. A later study, however, used visual recorded data of two different kinds. Cohen et al. (2002) used on the one hand an optical 3D motion tracking system (Optotrack) for recording speech dynamics and on the other hand a 3D laser scanner for creating an accurate (static) head model. The Optotrack system recorded the motion of 19 active infrared markers affixed to the speaker's face at 30 fps (left part of Figure 2.13, note the cables required for active IR markers). The laser scanner provided a 3D model of the speaker's face, which was then used to reshape their generic 3D head model to resemble the speaker, using manually specified correspondence points between the two head models (middle part of Figure 2.13). Together with a photograph-based texture map, this results in a 3D head more closely resembling the target speaker than the generic head model (right part of Figure 2.13). Starting from the control parameter sequence generated from the rule-based talking head, the parameter sequence was

[10] Example videos for UCSC's "Baldi" at `http://mambo.ucsc.edu/psl/international.html`.

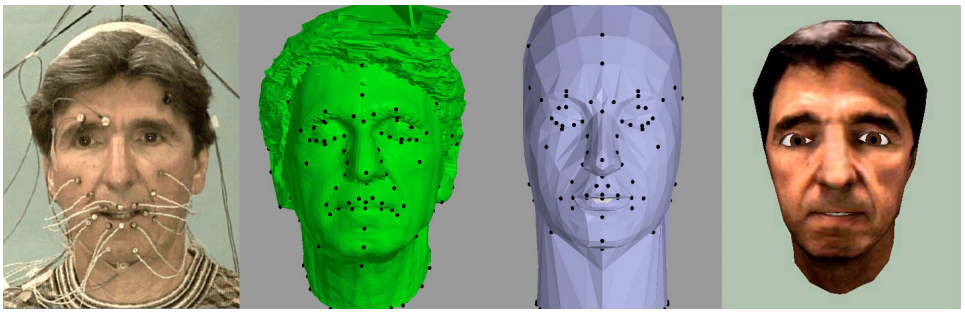


Figure 2.13: Facial motion tracking using active infrared markers (left), establishing correspondences between a 3D laser scan and the existing 3D head model (middle), final head model after morphing and texture mapping (images from Cohen et al., 2002).

automatically and iteratively refined such that the motion of the markers on the 3D head model became as similar as possible to the recorded marker motion.

Similarly, Jonas Beskow at KTH also initially worked with a rule-based system derived from Parke's work (Beskow, 1995; Beskow, 1997), then he collaborated closely with Cohen and Massaro of UCSC for some time (Cohen et al., 1998; Massaro et al., 1999) and later also turned more towards a measurement-based approach. Using simultaneous recordings of facial motion using a passive marker-based optical system (named Qualisys) and tongue motion using electromagnetic articulography (see Figure 2.14) of Swedish speech, Beskow et al. (2003) estimated trajectories for the control parameters of their head model which minimized the discrepancy between the motion induced by these parameters and the recorded marker motion, similar to Cohen et al. (2002). Control parameter trajectories produced in this way were then used by Beskow (2004) as training data for creating a motion model which generates parameter trajectories for given phoneme sequences. Several modeling techniques, among them dominance functions and artificial neural networks, were compared to each other as well as to an audio-only condition, and to the rule-based system. A subjective evaluation with 25 subjects showed that 1) the data-driven models did not differ significantly from each other, 2) the data-driven models achieved significantly better intelligibility than the audio-only condition, and 3) the intelligibility of the data-driven models was significantly *worse* than the rule-based system. This interesting last result is somewhat surprising. The study authors ascribe it to the fact that the rule-based system was particularly tailored to maximize clear articulation and intelligibility, and therefore tends to hyper-articulate (Beskow, 2004). Nevertheless, the results can be seen as a success for data-driven approaches. The same method was subsequently used for

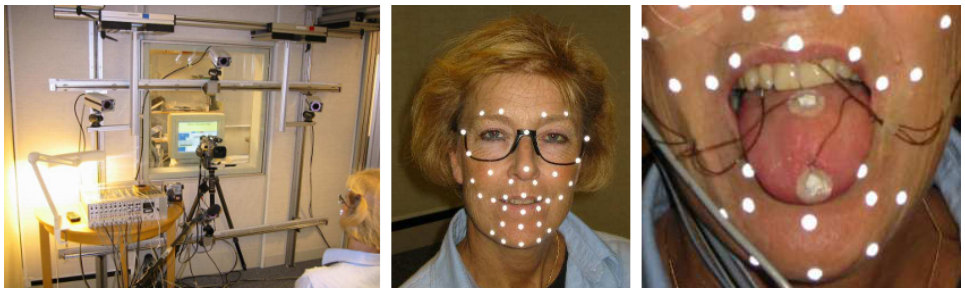


Figure 2.14: Hardware setup for face and tongue motion recording (left), passive optical markers on the speaker's face (middle) and active electromagnetic coils on the speaker's tongue (right). Images from Beskow (2003).

expressive visual speech for four different emotions (Beskow and Nordenberg, 2005), a task that would be very cumbersome for a purely rule-based system.[11]

A third group whose work is of high relevance to this dissertation is GIPSA-Lab in Grenoble, part of the French National Center for Scientific Research (CNRS) in partnership with several universities in Grenoble. After initially also following the route of dominance functions to control hand-crafted control parameters over time (Le Goff and Benoît, 1996), this group later approached the problem of audiovisual speech synthesis more from the direction of detailed recordings and careful analysis of speech production measurement data. Badin et al. (2000) (more detailed description in Badin et al., 2002) built a model of 3D motion and deformation for face and tongue based on magnetic resonance imaging and conventional video with green markers glued to the face and blue lip make-up (see Figure 2.15). They recorded 34 target articulation positions (central frames of key phones) from isolated word and vowel-consonant-vowel utterances, yielding 3D coordinates of 64 points, i.e., 34 shapes, each 192-dimensional. From this high-dimensional data, a low-dimensional set of parameters is determined via a "guided PCA" procedure: Applying PCA directly to the data would result in somewhat artificial components that are de-correlated and optimally explain the observed variance, but are difficult to interpret. Instead, Badin et al. (2000) iteratively choose arbitrary components (e.g., actually observed measures like jaw height), calculate a linear regression and subtract the motion which can thus be explained from the data corpus, and continue with the next component on the residual data, and so forth. The resulting components are more straightforward to interpret in terms of control, at the cost of suboptimal variance explanation and weak correlation between components (as opposed to no correlation). With six parameters, almost 97% of the variance

[11] Example videos for KTH talking heads at `http://www.speech.kth.se/august/august_eurospeech2.mpg` and `http://www.youtube.com/watch?v=X56XvZ_SBpw`.

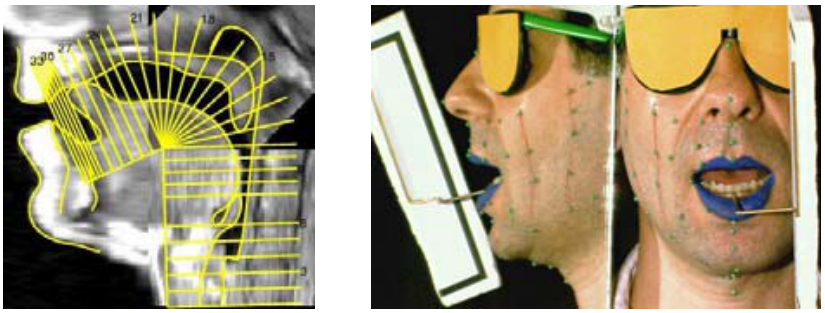


Figure 2.15: Articulation acquisition using magnetic resonance imaging (left) and conventional video (right). Images from Badin et al. (2000).

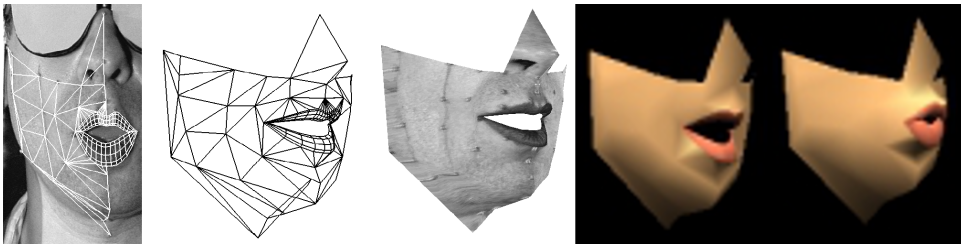


Figure 2.16: From left to right: Mesh resulting from marker positions overlaid onto a frame from the video recordings, the polygon mesh from a different viewpoint in 3D space, texture mapping and morphing applied to the mesh, two of the selected parameters in extreme positions (jaw opening and lip rounding). Images from Revéret et al. (2000).

within the key shapes could be explained. Using this kind of data, Revéret et al. (2000) built a text-to-audiovisual speech synthesis system where facial animation trajectories of the described components (obtained from guided PCA) are generated according to Öhman's numerical model of coarticulation (Öhman, 1967) using the phone targets and timings provided by a TTS system. Figure 2.16 illustrates the system of Revéret et al. (2000).[12] A similar system but with 168 recorded beads and thus a denser facial mesh was presented by Elisei et al. (2001), applying the procedure on several speakers in French and Arabic. Figure 2.17 illustrates this system.[13]

Soon after proposing their HMM-based speech parameter generation algorithm and the HTS speech synthesis framework (see also Section 2.1.2), the Tokyo/Nagoya group also applied their system to visual speech synthesis. Masuko et al. (1998) use the ability of the HTS system to synthesize smooth

[12]Example video at http://morpheo.inrialpes.fr/people/reveret/ttvs/ttvs_syn.avi.

[13]Example videos at http://www.isca-speech.org/archive_open/avsp01/av01_090.html.

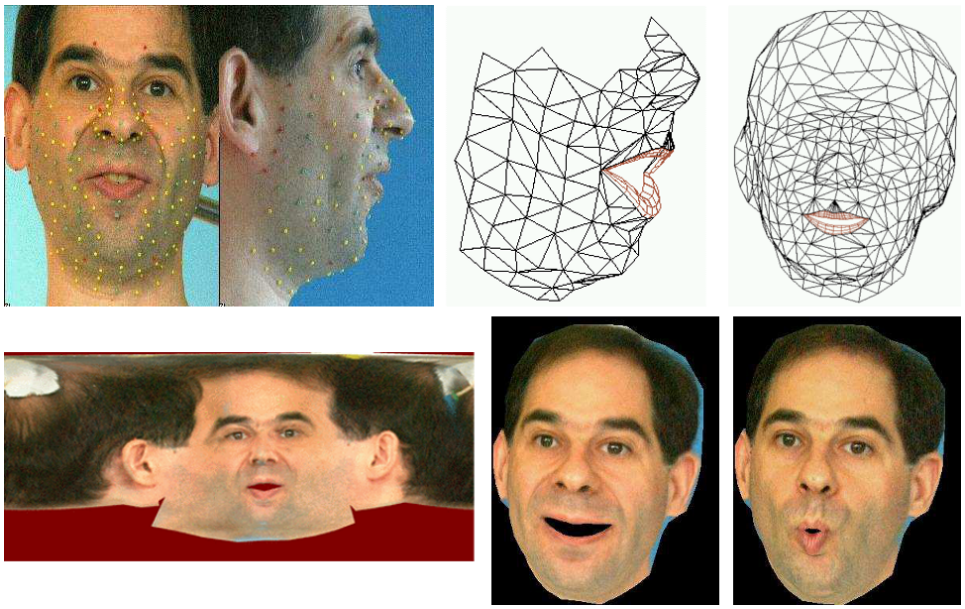


Figure 2.17: Acquisition of key face position with 168 beads attached to the speaker's face (top left). Facial mesh resulting from marker positions and extended mesh with additional rigid (non-deforming) parts (top right). Cylindrical texture created from photographs (bottom left). 3D head model with cylindrical texture applied (bottom right). Images from Elisei et al. (2001).

trajectories for lip motion parameters. From 216 phonetically balanced Japanese words video-recorded at 60 frames per second (after deinterlacing), they extract the inner lip contour by an automatic tracking procedure and manual corrections. Then the mouth position and shape is parametrized by the ten distance measures shown in Figure 2.18, plus their respective delta parameters (difference to preceding position) as dynamic features, i.e., a 20-dimensional observation vector for each video frame. Then 4-state syllable HMMs are trained on this data, from which new observation parameter sequences can be synthesized for any input text (text-to-visual-speech). An extension of this setup was presented by Tamura et al. (1998a), where joint audiovisual syllable HMMs are trained, using the lip parameters for the visual part and mel-cepstral coefficients for the acoustic part. These models are then used to produce lip movement for a given input speech signal: first, the audio part of the models is used to find the temporal syllable borders in the signal in a recognition step, and then lip motion trajectories are synthesized using these temporal borders (speech-driven visual speech). If the utterance of the input speech signal is known, the recognition step becomes easier and thus results in more accurate borders (text-and-speech-driven visual speech). With the same data, Tamura et al. (1999) also trained joint

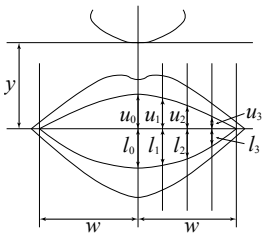



Figure 2.18: Ten parameters to characterize the inner lip contour (image from Masuko et al., 1998).

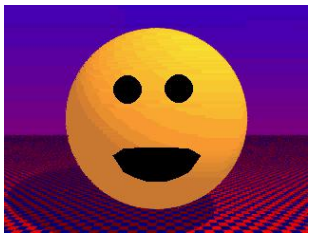


Figure 2.19: Video frame extracted from supplementary material published with Tamura et al. (1998a).

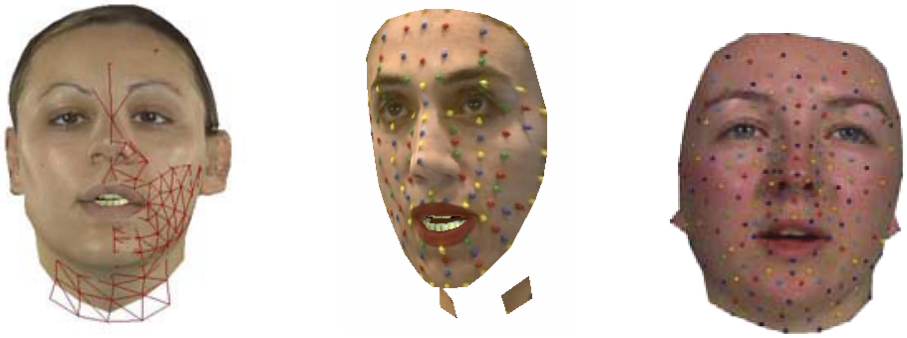


Figure 2.20: Photo-realistic textures for 3D head models (images from Govokhina et al., 2006, Bailly et al., 2008, and Bailly et al., 2009).

audiovisual HMMs to synthesize both an acoustic and a visual speech signal for given input text (text-to-audiovisual-speech). In this study, 7-state syllable HMMs and 4-state tri-phone HMMs were compared, which resulted in similar synthetic lip movement, but better acoustic speech from the tri-phone HMMs. Although this body of work may fall short to those discussed before in terms of capturing accuracy and visual detail (Figure 2.19 shows an example frame[14]), it is of high relevance because it is conceptually related to the system described in the following chapters of this dissertation.

GIPSA Lab in Grenoble have also made important contributions on audiovisual synthesis using HMMs. Govokhina et al. (2006) use an HMM-based trajectory formation system for articulatory gesture planning to guide a segment selection/concatenation procedure (similar to hybrid systems discussed in Section 2.1.3) for facial speech animation. In later studies, they synthesize facial motion trajectories directly from visual HMMs and add to each model a mean time lag to the temporal borders of the corresponding

[14]Example videos at http://www.isca-speech.org/archive_open/avsp98/av98_221.html.

Figure 2.21: Example frames generated by a hybrid HMM/unit selection system for speech-driven 3D facial animation (image from Tao et al., 2009).

acoustic phone unit (Govokhina et al., 2007), to improve synchronization. These mean time lags are learned iteratively using forced alignments of the training data. The systems of this group additionally also produce dynamic photo-realistic texture maps for the face models (Bailly et al., 2009), some examples can be seen in Figure 2.20.

Similar to the speech-driven facial motion synthesis described above (Tamura et al., 1998a), Hofer et al. (2008) use audiovisual HMMs for first aligning the trained models to a new input speech segment (recognition step using the acoustic parts of the models) in order to then synthesize 3D lip motion using the same unit borders (synthesis step using the visual parts of the models). In a follow-up study, Hofer and Richmond (2010) showed that an artificial neural network was able to outperform the HMM-based approach for this task.

A hybrid HMM/unit selection system for speech-driven 3D visual synthesis was proposed by Tao et al. (2009). They recorded several hundred utterances using 50 facial markers with synchronous audio. This data was then used to train audio and visual HMMs, which are then coupled into fused HMMs. The resulting models can then be used to drive the unit selection process (via target and concatenation costs) to find a marker motion sequence for given audio input. Their system is tailored to be very fast and thus suitable for real-time applications. A shortcoming of their study is that it does not include subjective evaluations; the quality of the resulting animations is assessed based on objective error measures and on the judgment of the authors. Figure 2.21 shows example frames created by their system.

Before concluding the subsection on model-based visual speech, it should be noted that the MPEG consortium developed the MPEG-4 Facial Animation standard (Pandzic and Forchheimer, 2003), which includes a set of facial landmarks (such as bottom of the chin), a set of facial animation

parameters (such as certain speech articulation gestures, or motion of specific landmarks) and a set of facial distance units (such as eye separation), which may be used as measuring units for specifying animation parameter amplitudes. Originally designed to allow very low bit rate compression and transmission of animation parameters, the MPEG-4 standard is used by some research groups in visual speech synthesis and not by others. It is not (yet) widely adopted in the animation industry. The research presented in this dissertation does not make use of MPEG-4, mainly because the facial motion capturing system we use does not follow the standard. Instead we used that system's format for simplicity.

### 2.2.3 Fusion of Image-based and 3D-based Paradigms

This section presented several approaches to visual speech synthesis, making a distinction into image-based and 3D-model-based methods. This border is becoming increasingly blurry as image-based methods start to incorporate some 3D features (e.g., L. Wang et al., 2011b, Figure 2.9) and as 3D methods start to use photo-realistic textures (e.g., Elisei et al., 2001, Figure 2.17). These efforts to increase realism are successful to a certain degree, but the problem of flexible scene illumination, for example, remains unsolved. Realistic shading requires a more detailed 3D mesh than just a rudimentary face shape to which the changing texture is applied, and one that actually deforms over time. And while it is true that a dynamic photo-realistic texture can provide additional detail (e.g., skin wrinkles) for a 3D model, these details are in fact not there from a shading point of view, which means they cannot correctly honor changes in illumination.

Recent advances in capturing and rendering technology make very high definition animated 3D faces (including wrinkles and even skin pores) possible, where the color information is simultaneously captured under controlled illumination conditions and thus suitable for "re-shading" in a differently lighted scene. Several companies have demonstrated impressive results in this area, e.g., Disney Research (Beeler et al., 2011)[15], DimensionalImaging[16] and Nvidia[17]; Figure 2.22 shows an example. Following such an approach, where deforming high-resolution facial meshes are recorded, results in extremely high-dimensional data. In order to use such data for synthesis, a suitable lower-dimensional representation needs to be found. Without having such data available to experiment with, it is difficult to say whether

[15] Disney Research example videos at http://graphics.ethz.ch/publications/papers/paperBee11.php.

[16] Dimensional Imaging example videos at http://www.youtube.com/watch?v=wriTh6pg7To and http://www.di3d.com/products/4d_systems/.

[17] Nvidia "FaceWorks" example videos at http://www.youtube.com/watch?v=STzAxVYUl4Y and http://www.youtube.com/watch?v=F9y-8IzNpQ4.

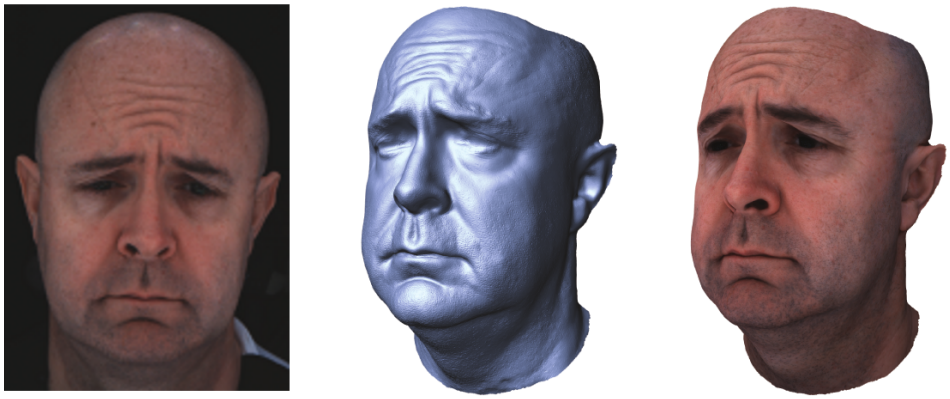


Figure 2.22: Image from one camera (left), untextured rendering of the 3D surface reconstructed from 7 cameras (middle), rendering with texture (right). Images from Beeler et al. (2011).

simple methods like PCA are sufficient or whether something more complex is required. Still, it seems possible that these recent advances in capturing and rendering technology may close the gap between image-based and 3D-based methods for visual speech synthesis, by providing 3D data (with all the flexibilities) and rendering techniques which achieve photo-realism.

## 2.3 Audiovisual Speech: Synchrony between Sound and Vision

A good audiovisual speech synthesis system, which produces acoustic speech and facial motion given some input text, does not only need to generate both kinds of signals in high quality; it also needs to achieve good temporal alignment between the two, i.e., synchrony between sound and vision, in order to deliver an audiovisual experience that is consistent with those delivered by a real speaking person.

For the intelligibility of speech in noise, it has been shown that simultaneous presentation of the speaking face has a significant beneficial effect (Sumby and Pollack, 1954), even when the speaker is a synthetic talking head (Ouni et al., 2007). This benefit is naturally dependent on correct alignment (within some tolerance interval) between speech and face motion. And even if intelligibility remains unaffected by small synchronization discrepancies, the perceived naturalness or overall quality may still be impaired.

A simple and very common synchronization approach, followed by the majority of the systems presented earlier, is to use the phone beginning and end times provided by a TTS system as input for visual speech synthesis.

In such a setup, the TTS system, responsible for producing the acoustic speech signal, and the visual speech synthesis system are treated as mostly independent, consecutive components. The TTS system processes the textual input, passes the sequence of phone symbols and their temporal borders (both of which can be seen as intermediate results in the process of producing the speech signal) on to the visual system, and continues to generate an acoustic speech signal. The visual system then produces a visual signal using the same temporal borders. For concatenative systems, this is a reasonable choice (e.g., if the acoustic and visual units are taken from an inventory which was recorded simultaneously, synchronization will always be correct anyways) and for rule-based systems there is maybe no alternative (since there is no visual data, the rules need to be defined such that they produce facial animations that "correctly" match the audio both in terms of movement and synchronization). However, for generative approaches (like HMMs) divergence between the two modalities can be a problem.

Tamura et al. (1999) address this problem by training joint audiovisual HMMs, which model the acoustic features and the mouth shape parameters of each frame together, as a combined, multi-modal observation, including dynamic features of both modalities. In their setup, the phone borders are trained as a common duration model for both modalities. From such a joint audiovisual voice model, it is straightforward to generate synchronous signals. The same kind of modeling is also used by the two-step systems of Hofer et al. (2008), Hofer and Richmond (2010), L. Wang et al. (2010) and L. Wang et al. (2011b), where the goal is to generate a visual signal for a given input audio signal. Govokhina et al. (2007) and Bailly et al. (2009) use an explicit phasing model to align the sequences generated by their acoustic and visual HMMs. The approach of joint audiovisual modeling is also applied in Chapter 6 (and Schabus et al., 2014a), where its benefit is analyzed in detail.

# Chapter 3

# Speech Synthesis Using Hidden Markov Models

This chapter is intended to provide an overview of statistical parametric speech synthesis using HMMs, in particular the parts that are relevant for extending an existing acoustic system to visual and audiovisual speech synthesis. Rather than starting from a fundamental introduction of HMMs, the more focused perspective here will follow the concrete procedure from the original speech data via feature extraction and training to the final synthesized speech. A more thorough introduction to HMMs can be found in the classical tutorial by Rabiner (1989). Overviews of speech synthesis using HMMs can also be found in the survey articles of Zen et al. (2009) and Tokuda et al. (2013), as well as in the PhD dissertations of Masuko (2002), Yoshimura (2002), Zen (2006), and Yamagishi (2006).

## 3.1 Audio Feature Extraction and Re-Synthesis

A recorded speech audio signal as a waveform, i.e., a one-dimensional signal at a very high sampling rate like 44 100 Hz (see Row 1 of Figure 3.1) is of limited use for speech processing. Both for human analysis of speech and for speech processing by machine, spectral information is of vital importance. Phoneticians typically use a three-dimensional graphical representation where the horizontal axis represents time, the vertical axis represents frequency, and the plotted color or intensity indicates the amplitude of the corresponding frequency at the corresponding point in time. This is called a spectrogram (see Row 2 of Figure 3.1), which allows straightforward visual identification of, for example, fricatives where high frequencies dominate (such as /f/ and /ʃ/ in Figure 3.1), as well as vowels with their distinct

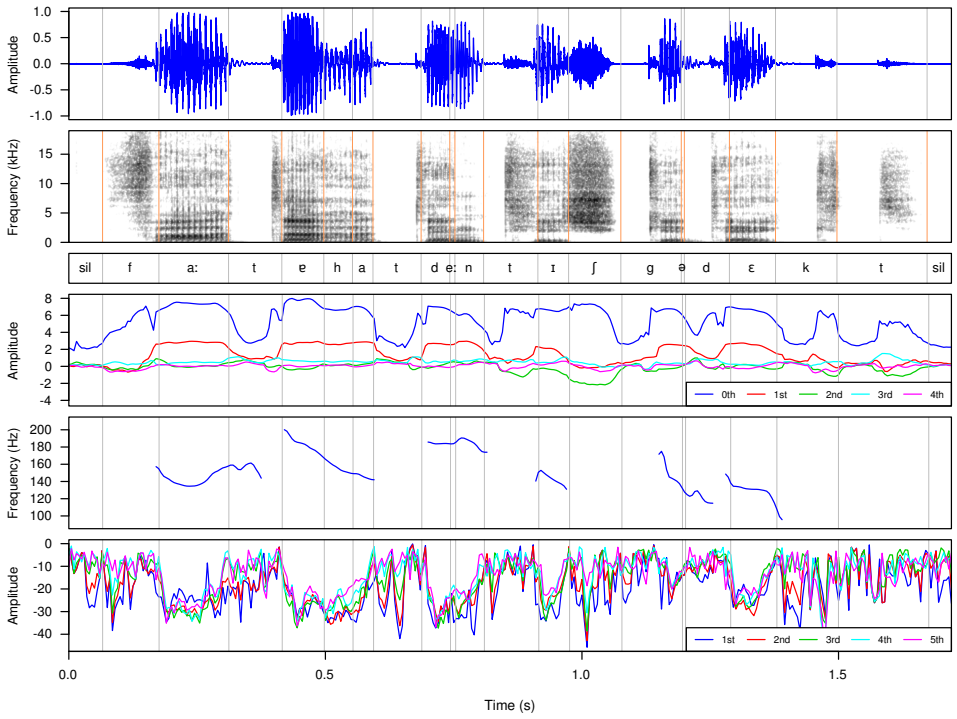



Figure 3.1: Audio features extracted from the German sentence "Vater hat den Tisch gedeckt." (father has set the table), phonetically /ˈfaːtɐ hat deːn tɪʃ gəˈdɛkt/, uttered by a male speaker (the author, in this case). Row 1: waveform of the recorded audio signal. Row 2: spectrogram. Row 3: automatically determined phone borders. Row 4: first 5 MFCCs. Row 5: fundamental frequency. Row 6: first 5 band-aperiodicity features.

formants (characteristic peaks in the frequency spectrum, e.g., dark horizontal bars in /aː/). Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) are typically used as acoustic features for spectral information in computational processing of speech and other audio signals. They are computed as follows.

To obtain the complex cepstrum of a speech signal segment, a complex Fourier transform is computed, followed by a complex logarithm, followed by an inverse Fourier transform (Oppenheim and Schafer, 1968). Typically, a frequency transformation is applied in this process, to obtain a frequency resolution similar to that of the human ear, e.g., a transformation to the mel-scale (Stevens et al., 1937) or a mel-generalized variant (Tokuda et al., 1994). The spectrum $H(e^{j\omega})$ may then be represented approximately by

$M + 1$ cepstral coefficients as

$$H(z) = \exp \sum_{m=0}^{M} c_m \cdot \left( \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \right)^m , \tag{3.1}$$

where $z = e^{j\omega}$ is any complex number with $|z| = 1$, $|\alpha| < 1$ is a real-valued constant which determines the frequency warping, and the number of coefficients $(M+1)$ determines the quality of the approximation. The coefficient vector $[c_0, c_1, ..., c_M]^\top$ is then used as the feature vector representing the spectral information of the speech signal segment (Fukada et al., 1992; Tokuda et al., 1994).

The human vocal tract acts as a dynamic spectral filter which allows the production of different speech sounds by changing the spectral properties over time. To capture these dynamic changes between sounds, the speech segments considered for spectral feature extraction need to be quite short, certainly shorter than the duration of a phone. A typical setup would be to use segments of 4096 samples, which corresponds to roughly 93 ms at a sampling rate of 44 100 Hz, with a frame shift of 5 ms, which results in strongly overlapping segments (0–93, 5–98, 10–103, etc.), to which a windowing function (e.g., a Blackman window) is applied and of which then the MFCCs are determined. A frame shift of 5 ms gives rise to 200 feature vectors per second (200 Hz), each of them typically of 40 dimensions ($M = 39$) for speech synthesis. Row 4 of Figure 3.1 shows the first five (of a total of 40) MFCCs extracted from the speech signal shown in Row 1. It can be seen that they are fairly smooth, and that their amplitude decreases with growing order.

For many applications, among them speech recognition and speaker verification, MFCCs can be sufficient as audio features. For speech synthesis, however, we need to be able to "invert" the analysis step to obtain a speech signal again, and for this an excitation signal is additionally required. In the simplest form, speech can be generated following a source-filter model, illustrated in Figure 3.2: the excitation source is chosen based on a voiced/unvoiced decision. For voiced speech sounds, a quasi-periodic train of pulses is generated according to the required pitch period, and for unvoiced speech sounds a sequence of random noise is generated. The excitation signal is then filtered by a slowly time-varying linear system, which uses spectral information as input (Imai, 1983; Fukada et al., 1992; Masuko, 2002). Therefore, the feature extraction procedure needs to extract pitch information in the form of fundamental frequency to allow re-synthesis. Row 5 of Figure 3.1 shows the fundamental frequency extracted from the speech signal shown in Row 1. Its value is defined only for segments corresponding to voiced speech sounds, and its frame rate is also governed by the frame shift, as for the MFCCs (a frame shift of 5 ms resulting in 200 Hz).

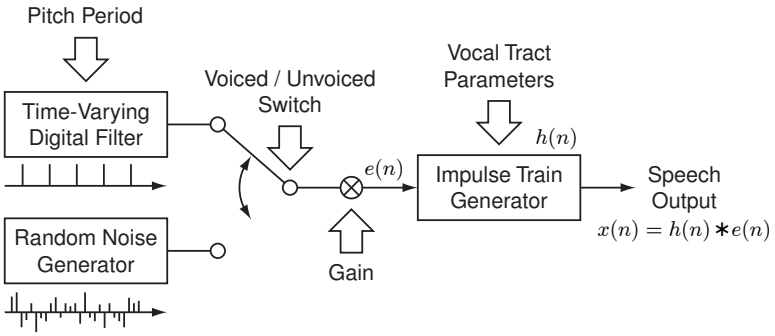



Figure 3.2: Discrete-time source-filter model for speech production (image from Masuko, 2002).

Furthermore, it is known that the vibration created by the vocal folds during speech is not perfectly periodic, and that incorporating such aperiodicities into voice coding can improve the achieved voice quality (Fujimura, 1968). A very common speech analysis and re-synthesis system called STRAIGHT (Kawahara et al., 1999; Kawahara et al., 2001) therefore also extracts aperiodicity features for several frequency bands and uses pitch-adaptive windowing for more accurate spectral feature extraction, in order to allow high-quality re-synthesis from the parameters. Row 6 of Figure 3.1 shows the first five (of a total of 25) band-aperiodicity features extracted from the speech signal shown in Row 1. Again, a 5 ms frame shift results in 200 parameter vectors per second. The MFCCs, the fundamental frequency and the aperiodicity features in Figure 3.1 were all extracted using the STRAIGHT analysis tools, and the audiovisual speech synthesis system presented in the following chapter also uses STRAIGHT for both analysis and re-synthesis. The HTS working group have adopted STRAIGHT for their speech synthesis system in 2005. Zen et al. (2007b) discuss the benefits of this analysis/re-synthesis system for HMM-based speech synthesis.

To summarize, an analysis/re-synthesis tool like STRAIGHT can be used on a recorded one-dimensional speech signal at 44 100 Hz to extract, e.g., 66-dimensional feature vectors ($40 + 1 + 25$ for mel-cepstral, fundamental frequency and band-aperiodicity features) at 200 Hz. Using the re-synthesis part of STRAIGHT, a speech signal can be generated from these parameters, yielding a one-dimensional signal at 44 100 Hz once again. Although of high quality, this speech encoding and decoding procedure is not perfect, in the sense that the speech signal re-synthesized from the parameters is not identical to the original recording. In particular, a certain "buzziness" of the voiced excitation is audible.[1]

[1] A listening example of the utterance of Figure 3.1 comparing recorded speech and speech re-synthesized from STRAIGHT parameters can be found at `http://schabus.xyz/phd/resynthesis`.

It should be noted that other speech parametrizations exist that are also used in HMM-based speech synthesis. Zen et al. (2009) provide an overview.

## 3.2 Phonetic Borders via Forced Alignment

We can use the described analysis procedure to extract a sequence of feature vectors for each recorded utterance. As described in the following section, we want to use phones as the fundamental modeling unit. Therefore, in order to use the extracted feature vector sequences of a collection of speech recordings as training data, a phonetic labeling of this data on the temporal axis is required. This can be done manually, typically by listening to the samples and looking at various graphical representations, like the spectrogram in Row 2 of Figure 3.1. However, this is an extremely time-consuming task, requires phonetic expertise, and poor inter- and intra-expert consistency can be a problem. For these reasons, automatic procedures are commonly used for this task (Leung and Zue, 1984), typically using HMMs in a fashion very similar to automatic speech recognition (Brugnara et al., 1993): Since the sequence of phones for a given sound file is already known, a collection of phone HMMs can be used to determine those temporal borders that maximize the likelihood of the observations (i.e., the feature vectors). This is sometimes called "forced alignment", because the phone sequence is given and fixed, and the aim is to find the best temporal alignment between the models corresponding to this phone sequence and the observed speech feature sequence. If available, the acoustic models of a general-purpose speaker-independent speech recognition system can be used for this procedure. Otherwise, an iterative "flat-start" approach can be used: At the beginning, each utterance of the training data is divided into phones equidistantly (because no better information is available). Then, acoustic models (e.g., HMMs modeling MFCCs) are trained on the speech segments according to this (poor) segmentation. Although the resulting models will not be very accurate, they can be used for forced alignment, resulting in a new segmentation of the data, which is very likely to be an improvement over the initial equidistant segmentation. This process is iterated several times, and it results in usable, albeit not perfect alignments.

The vertical lines in Figure 3.1 show the phonetic borders resulting from such a flat-start alignment procedure, with the respective phones given in Row 3. It can be seen that the resulting borders are generally reasonable: the initial silence, the high-frequency fricative /f/, the vowel /aː/ and the first /t/ consisting of a silent closure phase followed by a burst seem to be correctly aligned. On the other hand, the duration of the vowel /eː/ seems to be much too short. In this particular case, the reason for the suboptimal alignment stems from an assimilation carried out by the speaker: instead

of producing two plosives /t/ and /d/ in succession, only a single one was realized. So the procedure was "forced" to find an additional phone which is not actually there, resulting in this artifact, where the /d/ contains most of what should be part of the /eː/, and the latter becoming unnaturally short.

Despite the presence of such alignment errors, phonetic labelings created in this manner are usable in practice, and in fact all experiments presented in this dissertation are based on flat-start alignments, sometimes with small amounts of manual correction applied afterwards.

## 3.3 Training of Feature Models

Next, we want to train HMMs for the phonetically labeled features extracted from a collection of recorded utterances. The temporal modeling unit in speech recognition and synthesis is usually one phone. This can be said to be a natural choice, as it results in an inventory of manageable size (in contrast to, e.g., words) and it is linguistically well founded. Because we are interested in modeling concrete realizations, we stick to the term phone rather than phoneme here, but the difference between the two is small from an engineering perspective and the speech processing literature sometimes uses them interchangeably.

If we assume for now that one HMM is trained for each distinct phone appearing in our training data[2], and that, e.g., 41 phone symbols have been used to transcribe the utterances, then we need to use the observations present in the training data to train 41 phone HMMs such that they best "explain" the training data.

Figure 3.3 shows a typical HMM structure used for speech synthesis: a five-state left-to-right model with no skips and no self-loops. In contrast to HMMs in general, the state transition probabilities play no role in such a model, because all states are traversed sequentially from left to right, and the number of observations generated by each state is governed by explicit state duration Probability Density Functions (PDFs) instead of self-loop state transition probabilities, as discussed in more detail in the following Subsection 3.3.1. Therefore, the only parameters required to concretely define such a model are the five duration PDFs and the five output or observation PDFs; in both cases typically Gaussian distributions:

$$p_i(d) = \mathcal{N}(d \,|\, \mu_i, \sigma_i^2) \quad \text{for the duration of state } i, \tag{3.2}$$

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t \,|\, \boldsymbol{\mu}_i, \boldsymbol{U}_i) \quad \text{for the output of state } i, \tag{3.3}$$

[2] This is not actually the case, as discussed in Subsection 3.3.4.

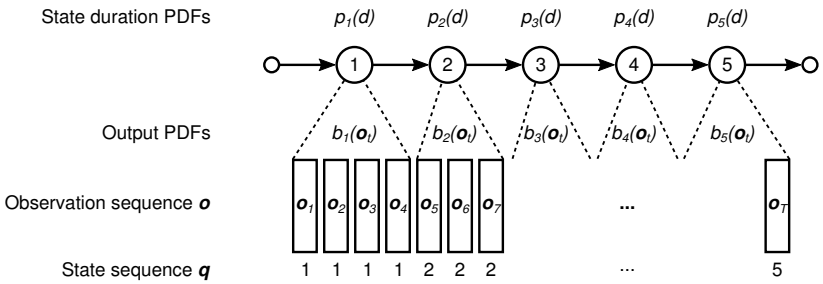



Figure 3.3: Structure of a typical HMM in speech synthesis: a five-state left-to-right model with no skips and with explicit state duration modeling (figure after Tokuda et al., 2013).

where $d \in \mathbb{R}$ is a random variable representing the number of observations, $\boldsymbol{o}_t \in \mathbb{R}^D$ is a $D$-dimensional random variable representing the observations, and $D$ is the dimensionality of the extracted speech feature vectors.

Such an HMM defines a stochastic generative process that generates feature vector sequences: According to the duration PDF of the first state $p_1(d)$, a concrete duration for state 1 is "drawn", i.e., the number of observations this state will need to generate ($d_1 = 4$ in Figure 3.3). Then, that many observation vectors $\boldsymbol{o}_t$ are "drawn" according to the state's output PDF $b_1(\boldsymbol{o}_t)$ (resulting in $\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3, \boldsymbol{o}_4$ in Figure 3.3). Likewise for the remaining states 2–5, resulting in the complete observation sequence $\boldsymbol{o} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ produced by the unobservable ("hidden") state sequence $\boldsymbol{q} = (q_1, q_2, ..., q_T)$.

Phone HMMs are concatenated to build utterance HMMs, both later on for synthesis, but also during training. At first, the phone HMMs are initialized using the data "associated" with each phone symbol via the given alignment, typically using the segmental $K$-means algorithm (Juang and Rabiner, 1990). Then, however, the HMM parameters are iteratively refined considering entire utterances, and the association of a certain observation vector to a certain state of a certain phone is no longer fixed but instead captured as a probability, defined by the HMM parameters. Given an observation vector sequence $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, ..., \boldsymbol{o}_T^\top]^\top$, the goal of the training procedure is to find the model $\lambda_{max}$ that is most likely to have generated the observation sequence $\boldsymbol{o}$:

$$\lambda_{max} = \operatorname*{argmax}_{\lambda} p(\boldsymbol{o}|\lambda) \tag{3.4}$$

$$= \operatorname*{argmax}_{\lambda} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q}|\lambda). \tag{3.5}$$

Note that the summation runs over all possible state sequences $\boldsymbol{q}$, whose number grows exponentially with the number of observations. Furthermore, there is no closed-form solution to this maximization problem. How-

ever, the efficient iterative re-estimation procedure of the Baum-Welch algorithm (L. E. Baum et al., 1970; Rabiner, 1989), an instance of the Expectation Maximization (EM) method (Dempster et al., 1977), is guaranteed to converge at least to a local maximum, and it is the usual method to train HMMs (not only in speech synthesis).

Once the models are trained, they can be used to generate an observation sequence $\boldsymbol{o}^*$ which is most likely to be produced by the respective utterance model $\lambda$,

$$\boldsymbol{o}^* = \underset{\boldsymbol{o}}{\operatorname{argmax}}\, P(\boldsymbol{o}|\lambda). \tag{3.6}$$

Synthesis of arbitrary utterances will be discussed in more detail in Section 3.4. First, the following four subsections discuss specific modeling concepts which are important for speech synthesis, namely explicit duration modeling, dynamic features, multi-space distributions and clustering of full-context models.

### 3.3.1 Explicit Duration Modeling: Hidden Semi-Markov Models

As already mentioned, the general way of modeling HMM state occupancy duration (i.e., the number of observations generated by a state) in terms of a self-loop transition probability is often avoided in speech synthesis. In a general HMM, the (implicit) duration probability density $p_i(d)$ of state $i$ with a self-transition coefficient $a_{ii}$ is

$$p_i(d) = (a_{ii})^{d-1}(1 - a_{ii}) \tag{3.7}$$

$$= \text{probability of } d \text{ consecutive observations in state } i, \tag{3.8}$$

and this exponential density is often inappropriate for modeling physical signals over time, such as speech (Rabiner, 1989). Instead, the duration is modeled by explicitly specifying a duration distribution for each state. After this modification, the model's stochastic process does not fulfill the Markov property anymore, which states that the conditional probability distribution of future states of the process depends only upon the present state. In the modified model, the next hidden state also depends on the amount of time spent in the current state, making it a so-called semi-Markov process. Accordingly, the modified model is called a Hidden Semi-Markov Model (HSMM) (a good overview is given by Yu, 2010).

Explicit durations have been proposed for HMM-based speech synthesis (and introduced to the HTS system) quite early by Yoshimura et al. (1998). However, the Gaussian distributions for each state were estimated from statistical variables obtained in the last iteration of the forward-backward algorithm.

This means that during the EM training of the HMMs, "traditional" models with self-transition probabilities were trained, which were discarded at the end, and for synthesis the explicit duration models were used. This inconsistency was cleared by Zen et al. (2004) and Zen et al. (2007d) by introducing the HSMM re-estimation formulae, where the PDFs for both state output and state duration are re-estimated during EM training, into the HMM-based speech synthesis framework (and also the HTS system).

The formulation by Zen et al. (2004) and Zen et al. (2007d) falls into the category of "explicit duration HMM", which Yu (2010) distinguishes from general HSMMs, "variable transition HMMs" and "residential time HMMs" in his classification of HSMM realizations. In general HSMMs, both the state and the duration are dependent on both the previous state and its duration. In contrast, for "explicit duration HMMs" the simplifying assumptions are made that the transition to the current state is independent from the duration of the previous state, and that the duration is only conditioned on the current state. Given the HSMM structure of Figure 3.3, which contains no branches and where the state transition probability between any two consecutive states is therefore equal to one, the state sequence is entirely determined by the mutually independent state duration distributions.

Although gamma distributions (Ishimatsu et al., 2001) and log-normal distributions (Yamagishi et al., 2004) have been applied to state duration modeling in HMM-based speech synthesis, Gaussian distributions are most commonly used, although it is known that they are not an optimal choice: The number of observations is inherently discrete and non-negative, the Gaussian distribution however is continuous and also negative values have a probability greater than zero. Nevertheless, simple one-dimensional Gaussian PDFs $p_i(d) = \mathcal{N}(d|\mu_i, \sigma_i^2)$ are often used to model the observed state durations $d_i$, and in the synthesis stage a simplified procedure is used in practice, where first all state durations (and thus the state sequence $\boldsymbol{q}$) are determined and then a speech feature vector sequence is generated for this fixed state sequence, rather than optimizing the likelihood across duration and observation PDFs simultaneously. The state durations $d_i$ are calculated as

$$d_i = \mu_i + \rho \cdot \sigma_i^2, \tag{3.9}$$

where $\mu_i$ and $\sigma_i^2$ are the mean and variance of the Gaussian distribution for the duration of state $i$, and $\rho$ is an acceleration parameter which can be used to control the speaking rate. "Normal" speaking rate is achieved with $\rho = 0$, faster speech with $\rho < 0$ and slower speech with $\rho > 0$.[3]

[3]Note however, that recent results from the FTW SALB project indicate that simple linear speaking rate scaling seems to be superior to this non-linear scaling based on the variance (Valentini-Botinhao et al., 2014).

### 3.3.2 Observation Modeling with Dynamic Features

In the stochastic generative process of the HSMM of Figure 3.3, the observation PDFs $b_i(\boldsymbol{o}_t)$ are responsible for generating the speech feature vector sequence for each state $i$. As described in Section 3.1, the speech signal is parametrized by mel-cepstral, fundamental frequency, and aperiodicity features, each at 200 Hz sampling rate. By regarding the three different feature vectors of one moment in time (with dimensionality 40, 1 and 25, e.g.) as one combined observation of dimensionality $D$ ($D = 40 + 1 + 25 = 66$), the features can be modeled simultaneously by a single $D$-dimensional observation PDF (Yoshimura et al., 1999). Typically, Gaussian distributions are used with the assumption of diagonal covariance matrices, i.e., the components are assumed to be independent (within and between the three different features). In this way, each of the $D$ components of the speech features is modeled by the (one-dimensional) mean and variance calculated from the values for this component appearing in all occurrences of the respective phone in the training data.

From the definition of the PDF for the $D$-dimensional Gaussian distribution (also called multivariate normal distribution) with mean vector $\boldsymbol{\mu}$ (of size $D \times 1$) and covariance matrix $\boldsymbol{\Sigma}$ (of size $D \times D$),

$$\mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{o}-\boldsymbol{\mu})\right), \tag{3.10}$$

it is obvious that $P(\boldsymbol{o}|\lambda)$ from Equation 3.6 is maximized when $\boldsymbol{o} = \boldsymbol{\mu}$, i.e., the mean vector is the most likely observation. Together with the assumption of conditional independence between the state output probability densities this entails that at each point in time, the most likely observation is the mean vector of the current state and hence that the generated observation sequence is a sequence of mean vectors. In the case of synthesizing speech signals from HSMMs, such unrealistic step sequences of features are problematic.

To overcome this problem, Tokuda et al. (1995) proposed to additionally include dynamic features into the observation vector. For a speech feature component $c_t$ for time $t$, the dynamic features $\Delta c_t$ and $\Delta^2 c_t$ are defined as

$$\Delta c_t = \sum_{i=-A}^{A} u_i \, c_{t+i}, \text{ and} \tag{3.11}$$

$$\Delta^2 c_t = \sum_{i=-B}^{B} v_i \, \Delta c_{t+i}. \tag{3.12}$$

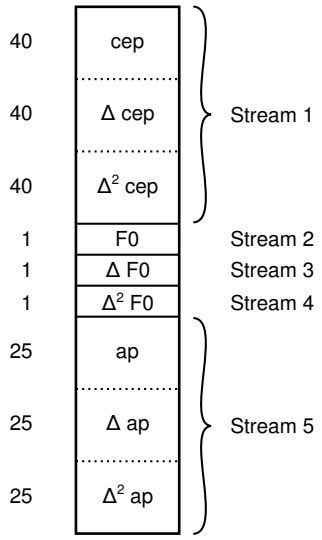



Figure 3.4: Structure of a typical 198-dimensional observation vector consisting of spectral, fundamental frequency and band-aperiodicity features and their respective dynamic features (figure after Yoshimura et al., 1999).

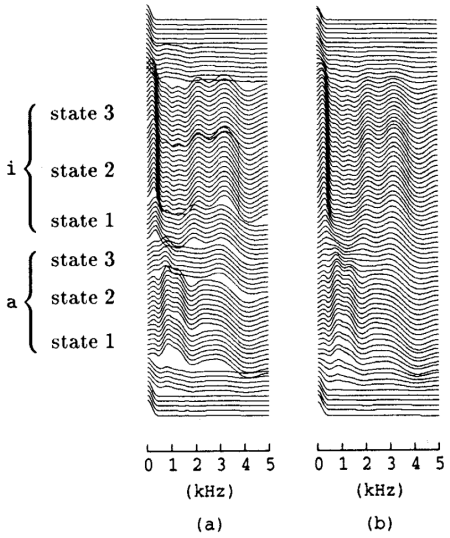



Figure 3.5: Example spectra generated from 3-state HMMs for the phone sequence sil, a, i, sil, without dynamic features (a) and with dynamic features (b). Image from Tokuda et al. (1995).

For example,

$$\Delta c_t = \frac{c_{t+1} - c_{t-1}}{2}, \text{ and} \tag{3.13}$$

$$\Delta^2 c_t = \frac{\Delta c_{t+1} - \Delta c_{t-1}}{2} = \frac{c_{t+2} - 2c_t + c_{t-2}}{4}, \tag{3.14}$$

i.e., the first and second order central difference quotients, which approximate the first and second order derivatives at time $t$ of the time-discrete feature trajectory. Adding the dynamic features triples the observation vector dimensionality (e.g., $(40 + 1 + 25) \cdot 3 = 198$), the resulting observation vector structure is illustrated in Figure 3.4.

Observation vectors augmented with dynamic features had previously been used successfully in speech recognition to improve the recognition rate (e.g., Furui, 1986), but the work of Tokuda et al. (1995) on speech synthesis contains an important innovation: the *generation* of parameters from an HMM under the constraints of the dynamic feature distributions in the maximum likelihood sense. Many concepts in HMM-based speech synthesis that remain important until today were already included in the work of Donovan and Woodland (1995a), Donovan and Woodland (1995b) and the PhD thesis of Donovan (1996), for example. However, the problem mentioned above

of piece-wise constant parameters per state was reported to be an issue by Donovan and Woodland (1995a), whereas the maximum likelihood parameter generation algorithm considering dynamic features by Tokuda et al. (1995)[4] successfully overcomes this problem, as illustrated in Figure 3.5: It is clearly visible that without dynamic features, all observations generated by the same state are identical, which creates discontinuities at the state transitions. With dynamic features, on the other hand, the spectrum varies smoothly over time. This may be seen as one of three key steps in making HMM-based speech synthesis popular, the second being the release and ongoing development of the open source system HTS (Zen et al., 2007a), and the third being the possibility of adaptation (discussed in Section 3.5).

Despite the undeniable success of using observation vectors that combine static and dynamic features, this approach may be seen as an ad-hoc engineering technique that works, but lacks theoretical elegance: Since the dynamic features are calculated from neighboring static features (cf. Equations 3.11–3.14), the relationship between them is completely deterministic. This fact is however ignored in the modeling, as the static and dynamic features are modeled as independent statistical variables, which allows inconsistencies between them. Zen et al. (2007c) have therefore proposed a reformulation which explicitly imposes the relationships between the static and dynamic features, resulting in a kind of model they named a "trajectory HMM", for which they also derived a Viterbi-type training algorithm. Although Zen et al. (2007c) reported successful improvements achieved in both speech recognition and speech synthesis, trajectory HMMs have not yet become widely adopted. They are also, for example, still considered a "recent development" by Tokuda et al. (2013).

### 3.3.3 Excitation Modeling Using Multi-Space PDFs

It was mentioned in Section 3.1 that the Fundamental Frequency (F0) is defined for voiced speech signal segments only, and this is also illustrated in the example of Figure 3.1 (Row 5). From the rough sketch of the re-synthesis procedure is is also clear that a voiced/unvoiced decision is required for each point in time. The F0 observation sequence thus needs to contain continuous one-dimensional values for voiced speech segments (i.e., feature extraction windows) and a discrete symbol that labels unvoiced speech segments as such. Neither a one-dimensional continuous probability density function nor a discrete probability distribution can model such observation sequences adequately.

For this reason, Tokuda et al. (1999) have proposed to use Multi-Space

[4] Interestingly, the two papers of Donovan and Woodland (1995a) and Tokuda et al. (1995) were presented at the same conference, namely ICASSP 1995 in Detroit.

Distributions (MSDs) to model F0 in their HMM-based speech synthesis framework. The idea of an MSD is that the observations come from a sample space $\Omega$ that is composed of $G$ subspaces $\Omega_1, ..., \Omega_G$:

$$\Omega = \bigcup_{g=1}^{G} \Omega_g. \tag{3.15}$$

Each subspace has a certain dimensionality $n_g$ and a certain PDF $\mathcal{P}_g(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^{n_g}$. Since they are PDFs, the $\mathcal{P}_g$ need to be properly normalized, i.e.,

$$\int_{\mathbb{R}^{n_g}} \mathcal{P}_g(\boldsymbol{x})\, d\boldsymbol{x} = 1. \tag{3.16}$$

Furthermore, there is a probability $w_g$ for each subspace, which indicates the probability that an observation is drawn from this subspace, i.e., $P(\Omega_g) = w_g$, where

$$\sum_{g=1}^{G} w_g = 1. \tag{3.17}$$

The probability distribution for the entire composed space can be expressed as

$$P(\Omega) = \sum_{g=1}^{G} P(\Omega_g) = \sum_{g=1}^{G} w_g \int_{\mathbb{R}^{n_g}} \mathcal{P}_g(\boldsymbol{x})\, dx = \sum_{g=1}^{G} w_g = 1. \tag{3.18}$$

The zero-dimensional case ($n_g = 0$) is used for discrete observations. It is assumed that the zero-dimensional space contains only one single point, and that $\mathcal{P}(\boldsymbol{x}) = 1$ for $n_g = 0$. An observation $\boldsymbol{o}$ of dimensionality $n$ is distributed according to

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g\, \mathcal{P}_g(\boldsymbol{x}), \tag{3.19}$$

where $S(\boldsymbol{o})$ is a set containing the indices of all subspaces that are of the same dimensionality $n$, i.e.,

$$S(\boldsymbol{o}) = \{g \,:\, n_g = n\} \;\text{ for } \boldsymbol{o} \in \mathbb{R}^n. \tag{3.20}$$

When all $n_g$ in an MSD are equal to zero, the resulting distribution is identical to the discrete distribution. When $G = 1$ and $n_1 = m \geq 1$, the resulting distribution is identical to an $m$-dimensional continuous distribution. When $G > 1$ and all $n_g$ are equal to the same value $m \geq 1$, the resulting distribution is identical to an $m$-dimensional $G$-mixture PDF. Hence, the MSD concept contains discrete, continuous and continuous mixture distributions as special cases.

Tokuda et al. (1999) and Masuko (2002) call HMMs that use MSDs as output distributions MSD-HMMs and give a comprehensive derivation of the re-estimation algorithm for MSD-HMM parameters in a general sense. In the particular case of F0 modeling, the setup they propose to apply is quite simple: the multi-space consists of a one-dimensional subspace for voiced regions of speech and a zero-dimensional subspace for unvoiced regions of speech (i.e., $G = 2$, $n_1 = 1$ and $n_2 = 0$), and the output probability of observation $o$ at state $i$ is

$$b_i(o) = \begin{cases} w_{i,1}\,\mathcal{N}(o|\mu_i, \sigma_i^2) & \text{(voiced)} \\ w_{i,2} & \text{(unvoiced)}, \end{cases} \tag{3.21}$$

where $w_{i,1}$ and $w_{i,2}$ are the probabilities of the observation being voiced or unvoiced, respectively, and $\mathcal{N}(o|\mu_i, \sigma_i^2)$ is a one-dimensional Gaussian distribution for the F0 values.

As discussed in the previous section, neighboring observations are used to calculate the dynamic features at each observation (cf. Equations 3.11–3.14). This poses a problem at the point of change from unvoiced to voiced (or vice-versa), because not all values to calculate $\Delta$F0 and/or $\Delta^2$F0 are defined (more precisely, they do not all come from the same subspace), even though the static F0 value is indeed defined. Therefore, the dynamic features are treated as unvoiced if any of the values required for calculating them is unvoiced. As a consequence, the change from voiced to unvoiced (or vice-versa) does not happen at the same point in time (i.e., at the same observation) for F0, $\Delta$F0 and $\Delta^2$F0, and therefore they are modeled as three independent streams in the HTS system, each with an MSD distribution as in Equation 3.21. Figure 3.4 from the previous section illustrates the five streams of the feature vector.

### 3.3.4 Full-context Modeling and Decision-Tree-Based Context Clustering

At the beginning of Section 3.3, where feature model training was introduced, the assumption was made that one HSMM would be trained per distinct phone symbol appearing in the training data, e.g., 41 HSMMs for a training corpus that uses a phone set of size 41. However, such a system design would be a poor choice: It is well known that the way a certain phone is realized is strongly influenced by contextual factors, e.g., neighboring phones. The tongue and other articulators do not jump from one position to the next but move in a continuous and often "optimizing" fashion. For example, the /t/ phones in the two words "tea" and "tree" are quite different, in this case due to an anticipatory assimilation effect (retracted

and somewhat labialized /t/ before /ɹ/).[5] One option to adequately model such co-articulation effects would be to use a larger, more fine-grained phone set. Another, more elegant option is to apply context-dependent modeling and let the training procedure handle the problem.

Bahl et al. (1980) introduced the idea of context-dependent phone modeling for automatic speech recognition, and Bahl et al. (1991) subsequently used binary decision trees for clustering the resulting models. The PhD dissertation of Odell (1995) provides an elaborate treatment of this subject and also introduced these techniques into the HTK speech recognition toolkit.

The general idea is easiest explained by a concrete example. Let us consider the utterance "There was a change now" and let us assume the utterance is represented by the following phone symbol sequence:

```
(dh, eh, r, w, aa, z, ax, ch, ey, n, jh, n, aw)
```

In a context-*independent* setup, we would simply use the symbol `ey` as the label for the speech feature vector sequence corresponding to the diphthong in the word "change". In a context-*dependent* setup, on the other hand, we use for the same feature vector sequence a label which is composed of that symbol plus its context, for example, `ch-ey-n` (a so-called tri-phone) or `ax-ch-ey-n-jh` (a so-called quin-phone). This allows for more accurate modeling, because the variation among all sequences bearing the same label is greatly reduced. Of course, the number of distinct labels, and hence HSMMs to train, becomes much larger, and the number of training instances per label becomes smaller, for a given data corpus.

For example, the well-known American English CMU Arctic SLT speech data corpus[6] consists of 1132 utterances and a total of 38 866 phone instances. It contains 41 distinct phones, 9546 distinct tri-phones and 28 662 distinct quin-phones. This means that for many quin-phones, there is only a single training instance available, which is obviously insufficient for robust HSMM parameter estimation. It furthermore means that most of the theoretically possible $41^5 \approx 116$ million quin-phones do not appear at all in the training data.

To overcome these problems, similar contexts need to be grouped together such that for each group there is a sufficient amount of training data, where "similar" should mean similarity in terms of realization, i.e., in terms of the data in the feature vector sequences (rather than similarity of the labels). The most successful method for this grouping is to cluster the contexts using a binary decision tree, where the inner nodes of the tree are yes/no questions

[5] This holds for British Received Pronunciation as well as General American pronunciation, but not necessarily for other varieties of English.

[6] `http://www.festvox.org/cmu_arctic/index.html`

about the phonetic context, and the leaves of the tree define the groups or clusters. Odell (1995) contrasts this top-down approach with bottom-up clustering of similar models and lists the following advantages from a speech recognition perspective:

- Due to the hierarchical structure, all models are "equally context-dependent", there is no need for less specific back-off models for contexts that have not occurred in the training data.
- Expert knowledge can be incorporated into the system by defining appropriate phonetic/linguistic questions which are used in the decision tree.
- During the construction of the tree, we can ensure that a sufficient amount of training examples is available by stopping the splitting procedure when a certain threshold is reached.
- Arbitrarily detailed additional contextual questions can be added, without the risk of underrepresented leaf nodes, if the tree is constructed carefully.

In speech recognition, unseen contexts as mentioned in the first point might come from a language model and a pronunciation dictionary which are used to build the model trellis of all possible utterances before recognition. For speech synthesis, a TTS system needs to be able to produce an appropriate speech signal for any given input context, also unseen ones.

To construct a binary clustering tree, a set of phonetic questions is required, which can be answered affirmatively or negatively for a given speech segment based on the information in its label. Because the clustering procedure is completely automatic, even more fine-grained criteria than "just" the five phones of the quin-phone can be used. As a typical example for speech synthesis, the question set included in the American English HTS demo package,[7] which uses the aforementioned CMU Arctic SLT corpus, is described in the following. All given examples are such that the answer to the question is yes for the example from the beginning of this subsection (`ey` from the sentence "There was a change now").

For each phone symbol $\phi$ in the phone set of the corpus (41 phones), the question set contains the following five questions concerning the five phones of the respective quin-phone:

- Is the current (`C`) phone $\phi$? Example: `C-ey`
- Is the phone preceding the current phone (left, `L`) $\phi$? Example: `L-ch`
- Is the phone before that (left-left, `LL`) $\phi$? Example: `LL-ax`

[7] http://hts.sp.nitech.ac.jp/?Download

- Is the phone succeeding the current phone (right, `R`) $\phi$? Example: `R-n`
- Is the phone after that (right-right, `RR`) $\phi$? Example: `RR-jh`

Furthermore, there are questions related to 60 (overlapping) phone classes (e.g., vowels, consonants, stops, nasals, fricatives, rounded vowels, unrounded vowels, voiced fricatives, unvoiced fricatives, etc.). The question set contains five questions for each phone class $\Phi$ following the same pattern LL-$\Phi$, L-$\Phi$, C-$\Phi$, R-$\Phi$, RR-$\Phi$ as for the phone identity questions above.
Examples: `LL-Back_Vowel`, `L-Affricate_Consonant`, `C-Diphthong_Vowel`, `R-Nasal`, `RR-Voiced_Fricative`.

There are also several positional questions included in the question set, following the pattern: Is the position of the current {phone, syllable, word} in the current {syllable, word, phrase}, relative to its {beginning, end} {equal to, less than, more than} {1, 2, 3, 4, 5, 6, 7}?
Examples: `Pos_C-Phone_in_C-Syl(Fw)==2`, `Pos_C-Word_in_C-Phrase(Bw)<=3`

Similarly, there are questions concerning various counts, like the number of phones in the {previous, current, next} syllable, the number of syllables in the {previous, current, next} word etc.
Examples: `R-Word_Num_Syls==1`, `Num-Syls_in_Utterance<=10`

There is furthermore a number of questions regarding (lexical) stress and (intonational) accent of syllables, including whether or not the {previous, current, next} syllable is stressed/accented or not, and also several questions addressing positions and counts regarding stressed/accented syllables.

More complete lists are given in Zen et al. (2007a) and Tokuda et al. (2013); an exhaustive list can be found in the questions file in the HTS demo package mentioned above, in which a total of 1483 questions are defined. For answering all these questions for a given feature vector sequence, the respective information needs to be present in the label for that sequence. Therefore, in practice the labels do not only contain the five phones of the quin-phone, but around 50 additional fields giving the various positions, counts, etc. required to answer all questions. Such feature-rich contextual descriptions are referred to as full-context labels.

It is worth noting that many more features may be added to the question set (and the full-context labels), depending, e.g., on the language, dialect/accent, speaking style, emotion, application domain, etc. In general, any split into two disjoint sets (corresponding to all contexts for which the question is answered affirmatively and negatively, resp.) which can be made on the symbolic level (i.e., based on the information in the labels) and which can be expected to discriminate the two sets on the signal level (based on some similarity measure for feature vector distributions) makes a good candidate for a question.

Next we consider how the clustering trees are built. The procedure starts with a collection of initial models, one for each full-context label appearing in the training data. In the case of the CMU Arctic SLT corpus and the question set described above, the entire corpus of 38 866 phones contains 38 658 distinct full-context labels (i.e., almost all full-contexts appear only once). For each of these a feature vector distribution has already been estimated based on the available training data. At the beginning, all models are in one big cluster. From all questions in the question set that have not yet been used, the one question that "best" splits the models into two clusters is selected. This question (e.g., `C-Vowel`) becomes the root node of the decision tree, with two child nodes corresponding to the clusters containing all models for which the answer to the question was yes (e.g., vowels) and no (e.g., non-vowels), respectively. This process is iterated, by repeatedly splitting all leave nodes (clusters) into two using the respective "best" question, until some stopping criterion is met.

Both open issues, i.e., how to determine the optimal question for the next split and when to stop further splitting leaf nodes, can be addressed by the Minimum Description Length (MDL) criterion (Rissanen, 1978). Given a set of models $\{1, ..., i, ..., I\}$, the description length $l_i(x^N)$ of model $i$ for the data collection $\{x^N = x_1, ..., x_N\}$ is defined as

$$l_i(x^N) = -\log P_{\hat{\theta}^{(i)}}(x^N) \; + \; \frac{\alpha_i}{2} \log N \; + \; \log I, \tag{3.22}$$

where $\alpha_i$ is the number of free parameters of model $i$ and $\hat{\theta}^{(i)}$ is the collection of maximum likelihood estimates for these parameters, $\hat{\theta}^{(i)} = (\theta_1^{(i)}, ..., \theta_{\alpha_i}^{(i)})$. The first term is the negative log-likelihood for the data, whose value decreases with increasing model complexity. The second term can be interpreted as a penalty for high model complexity. The third term is the code length required for choosing model $i$ and may be assumed to be constant. Hence, the model with minimal description length will be the one with the best trade-off between data explanation accuracy and model size (thus following an intuition similar to the principle of "Occam's razor" (Domingos, 1999)).

Invented by Rissanen (1978), the MDL criterion was applied successfully to tree-based clustering of continuous-density HMMs for speech recognition by Shinoda and Watanabe (2000) using several simplifying assumptions, and subsequently to speech synthesis in a similar fashion by Yoshimura (2002). MDL-based clustering has been available in the HTS system since version 1.0 (December 2002).

Each of the three speech features (spectral, fundamental frequency and aperiodicity) and also state duration may be influenced by different contextual factors, and thus their distributions are not pooled together but separate

Table 3.1: Number of leaf nodes in the clustering trees after training an American English voice using the CMU Arctic SLT data

| | State | | | | | |
|---|---|---|---|---|---|---|
| Feature | 1 | 2 | 3 | 4 | 5 | Total |
| Mel-cepstral | 275 | 238 | 235 | 260 | 275 | 1265 |
| Log F0 | 518 | 724 | 944 | 936 | 633 | 3755 |
| Duration | | | | | | 562 |

clustering trees are constructed per feature. Furthermore, the feature vector sequences represented by each of the individual states of a full-context HSMM are quite different (beginning, middle and end of the respective phone), and thus the distributions per HSMM state are also clustered independently for the speech features. For duration, the five one-dimensional distributions corresponding to the number of observations for each of the five states are actually combined into one five-dimensional distribution, resulting in only one single clustering tree for duration. Therefore, a total of $5 \cdot 3 + 1 = 16$ clustering trees are built in a setup with five-state HSMMs and the three features discussed earlier in this chapter.

Table 3.1 gives the number of leaf nodes in all clustering trees (and hence the number of estimated distributions) after the HTS demo training using the CMU Arctic SLT data has finished. In this case the demo training using the mel-generalized cepstral vocoder (Tokuda et al., 1994) (i.e., not STRAIGHT) was used, therefore there are no aperiodicity features. It can be seen that even for the most diverse group of distributions (state 3 of log F0), the originally 38 658 distributions have been grouped into 944 clusters, a reduction of more than 97 percent.

Figure 3.6 shows part of the MFCCs clustering tree for the second HSMM state. Firstly, this illustrates how the splitting procedure started with rather general questions concerning the current phone (Is it a vowel? If no, is it a voiced consonant? etc.) and that questions regarding the neighboring phones (e.g., `L-No_Continuent`) or position (e.g., `Seg_Bw<=1`) only begin to appear at a certain depth. Secondly, we can use the part shown here (in which only three of the 238 leaves are shown, cf. Table 3.1) to see which distribution would be used for our example context `ey` from the sentence "There was a change now" by following the following path from the root to a leaf node:

- Is the current phone a vowel? Yes, `ey` is a vowel.
- Is the vowel of the current syllable a front vowel? Yes, the `ey` itself is the vowel of the single-syllable word "change", and it is a front vowel.

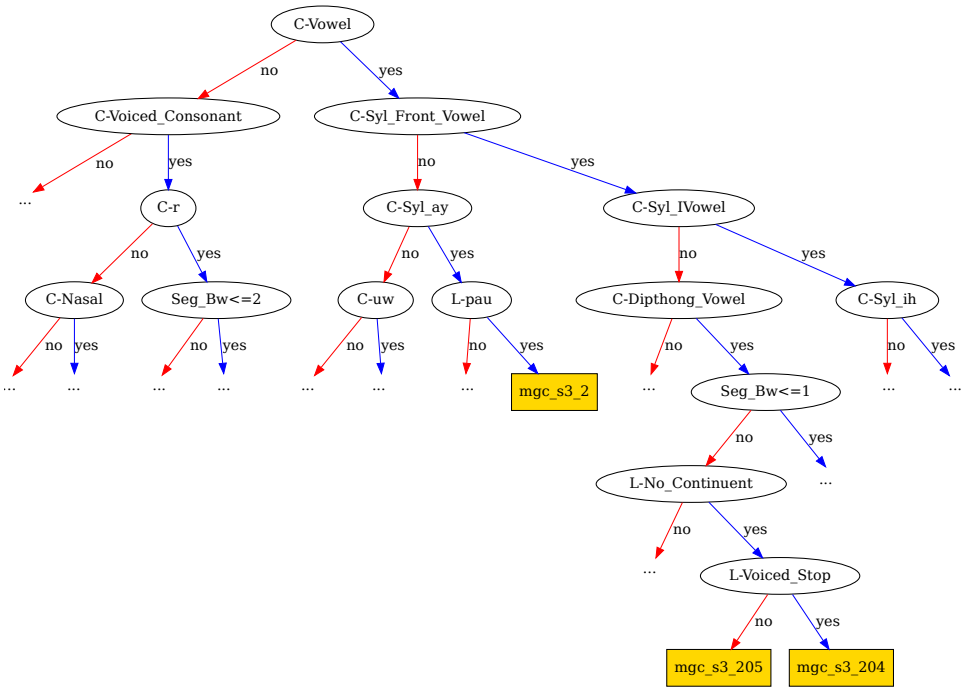



Figure 3.6: Part of the clustering tree for the spectral features of the second HSMM state after running the HTS demo training on the CMU Arctic SLT data.

- Does the vowel of the current syllable belong to the group of i-like vowels? No, `ey` is not an i-vowel.
- Is the current phone a diphthong? Yes, `ey` is a diphthong.
- Is the current phone the last phone of the syllable it appears in? No, the `ey` is followed by a `n` and a `jh`.
- Is the left neighbor phone not a continuent? Yes, the `ch` is a non-continuent.
- Is the left neighbor phone a voiced stop? No, `ch` is neither voiced nor a stop (it is an affricate).

The path ends in a leaf node containing the identifier of a distribution, `mgc_s3_205`. This tells us that this distribution, which models spectral features of the second HSMM state, was trained using data from the feature vector sequence extracted from the `ey` of the sentence "There was a change now", which was part of the training data. Furthermore, it tells us that if we were to synthesize such a context, we should use this distribution for the second state's spectral features.

This completes the section on model training. In practice, several of the steps are iteratively repeated to obtain more robust estimates, but all the important concepts have been discussed in this section.

## 3.4 Synthesis of Arbitrary Utterances

This section describes how we synthesize speech for an arbitrary input text, given the end result of the training procedure, i.e., a collection of clustering trees and the distributions associated to all leaf nodes.

As described in Chapter 2, the input text is first processed by a transcription stage, which typically employs a normalization component that converts abbreviations, numbers, etc. into their written-out form; a pronunciation dictionary where the phone sequence, syllable borders, lexical stress and part-of-speech for the words of the input text can be looked up; and prediction modules for the aforementioned, for words that are not found in the dictionary. The output of the transcription stage is a sequence of full-context labels, which contain all the symbolic information required to correctly traverse the clustering decision trees.

Next, a single utterance HSMM is constructed by concatenating phone HSMMs. Using for each of these the model structure of Figure 3.3 (five states, each with their duration and observation distributions) and the observation vector structure of Figure 3.4 (40-dimensional spectral, 1-dimensional fundamental frequency, and 25-dimensional aperiodicity features, each with their delta and delta-delta features), we need to obtain all the correct distributions: At each of the five states, a distribution for the duration, one for the spectral features, one for the fundamental frequency and one for the aperiodicity features are required. By traversing each of the 16 decision trees using the full-context label generated by the transcription stage for the respective phone, 16 leaf nodes and their associated distributions are found, and the three feature distributions at each state are stacked to form five observation distributions. Finally, we obtain an utterance HSMM consisting of $n$ states (where $n = 5k$ if $k$ is the number of phones), each with a one-dimensional duration distribution and a 198-dimensional observation distribution.

After building the utterance HSMM $\lambda$, the next step is to compute the speech feature vector sequence that model $\lambda$ is most likely to generate. Using the assumption of independence between state durations and state output as discussed in Section 3.3.1, we can already determine the state sequence using the duration distributions alone. For "normal" speaking rate ($\rho = 0$) the variances can be disregarded (cf. Equation 3.9) and the number of observations generated by each state is simply the (rounded) mean of the

corresponding duration distribution, resulting in the state sequence

$$\boldsymbol{q} = (q_1, ..., q_T) = (\underbrace{s_1, s_1, ..., s_1}_{d_1}, \underbrace{s_2, s_2, ..., s_2}_{d_2}, \ ..., \ \underbrace{s_n, s_n, ..., s_n}_{d_n}), \tag{3.23}$$

where the $s_i$ are the identifiers of the $n$ states of the utterance HSMM and the $d_i$ are the respective numbers of observations.

As discussed in Section 3.3.2, the observations $\boldsymbol{o}_t$ modeled by the HSMMs are composed of the "original" speech feature observations $\boldsymbol{c}_t$ and their dynamic and acceleration features, i.e.,

$$\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top]^\top. \tag{3.24}$$

But we are not interested in actually generating sequences for the dynamic features; Rather, the goal is to generate a sequence of static features which honors (in the maximum likelihood sense) the distributions for the static as well as those for the dynamic features. The relationship between the sequences $\boldsymbol{c} = [\boldsymbol{c}_1^\top, ..., \boldsymbol{c}_T^\top]^\top$ and $\boldsymbol{o} = [\boldsymbol{o}_1^\top, ..., \boldsymbol{o}_T^\top]^\top$ can be expressed in matrix form as

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}. \tag{3.25}$$

Using this relationship, the static observation sequence $\boldsymbol{c}^*$ that the utterance HSMM $\lambda$ is most likely to generate can be found from the state output distributions $P(\boldsymbol{o}|\lambda)$ as

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P(\boldsymbol{W}\boldsymbol{c}|\lambda) = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, \mathcal{N}(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}). \tag{3.26}$$

By equating the partial derivative of the logarithm of Equation 3.26 with respect to $\boldsymbol{c}$ to $\boldsymbol{0}$, a set of linear equations can be obtained to solve for $\boldsymbol{c}^*$:

$$\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\boldsymbol{c}^* = \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}, \tag{3.27}$$

$$\boldsymbol{c}^* = (\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}. \tag{3.28}$$

By exploiting the special structure of $\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}$, Equation 3.27 can be solved very efficiently using the parameter generation algorithm by Tokuda et al. (1995), or by using Cholesky decomposition or QR decomposition (Tokuda et al., 2000). A concise explanation of the parameter generation, including a visual explanation of Equation 3.25 is given by Zen et al. (2007c) (also found in Zen, 2006).

After $\boldsymbol{c}^*$ has been found, which is a sequence of, e.g., $40 + 1 + 25 = 66$-dimensional vectors, we can decompose it into the vector sequences of the individual speech features (spectral, fundamental frequency and aperiodicity

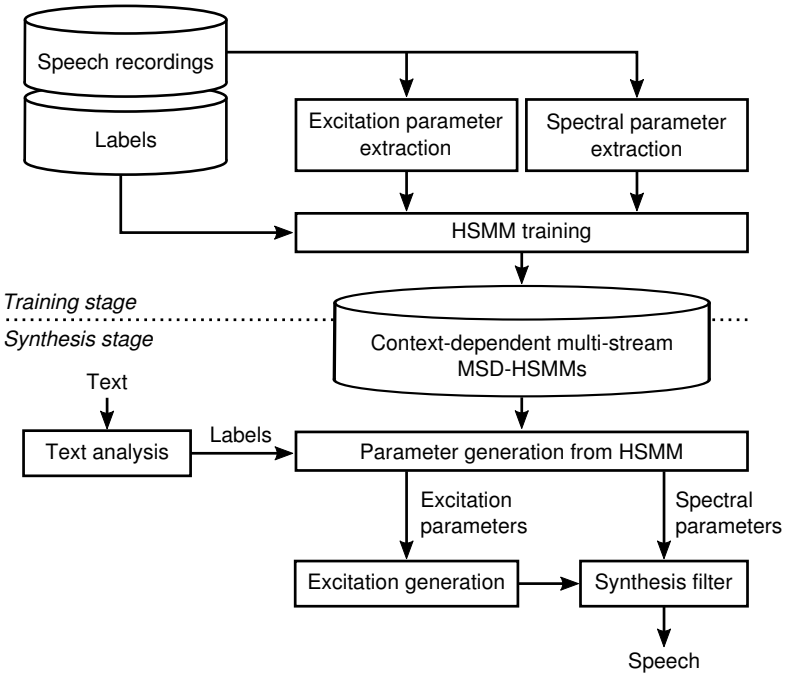



Figure 3.7: Schematic overview of an HSMM-based speech synthesis system (figure after Yoshimura et al., 2001).

features), and finally use the re-synthesis procedure (cf. Section 3.1) to create from those the final speech waveform. Figure 3.7 summarizes the HSMM-based speech synthesis system: In the training stage, parameter vectors are extracted from a collection of speech recordings. These parameters are then used together with the phonetic/temporal labels to train HSMMs. In the synthesis stage, a transcription module translates the input text into a sequence of phonetic labels. The most likely parameter vector sequence for this label sequence is generated from the HSMMs, and the parameters are re-synthesized to a speech signal.

## 3.5 Average Voices and Adaptation

An important advantage of the statistical parametric approach to speech synthesis—in particular in contrast to concatenative methods—is the ability to train average voice models across multiple speakers and to adapt such average voice models towards a new target speaker. Like the concept of speech modeling with HMMs itself, also the idea of adaptation originated in the speech recognition field (Gauvain and Lee, 1994; Digalakis et al., 1995; Leggetter and Woodland, 1995).

The most straightforward use case for adaptation in both speech recognition and speech synthesis is to obtain a high quality model for a specific target speaker, for whom only a small amount of speech data is available, by adapting from an average voice model that was previously trained using a large multi-speaker speech database (e.g., Tamura et al., 1998b). The advantages are evident: As soon as a large, high-quality average voice model is available, many different speakers' voices can be created with a small recording and training effort required for each of them. Figure 3.8 shows a schematic overview of a speaker-adaptive HSMM-based speech synthesis system. The synthesis stage is the same as in Figure 3.7; however, the training here now consists of two parts. First, an average voice is trained using a multi-speaker database. Then, the resulting average voice models are adapted using a target speaker database.

A multi-speaker training data collection for an average voice includes many speaker-dependent characteristics. However, the goal is to adapt to a wide variety of target speakers, therefore the average voice model should avoid such speaker-dependent characteristics. To overcome this problem, Yamagishi and Kobayashi (2007) proposed a Speaker-Adaptive Training (SAT) algorithm, which normalizes the influence of speaker differences by means of Maximum Likelihood Linear Regression (MLLR). The difference between each of the speakers and the canonical average voice is expressed by a simple linear regression function applied to the means of both the observation and the state duration distribution:

$$\boldsymbol{\mu}_i^{(f)} = \boldsymbol{\zeta}^{(f)}\boldsymbol{\mu}_i + \boldsymbol{\epsilon}^{(f)} = \boldsymbol{W}^{(f)}\boldsymbol{\xi}_i, \text{ and} \tag{3.29}$$

$$m_i^{(f)} = \chi^{(f)} m_i + \nu^{(f)} = \boldsymbol{X}^{(f)}\boldsymbol{\phi}_i. \tag{3.30}$$

where $\boldsymbol{\mu}_i^{(f)}$ and $m_i^{(f)}$ are respectively the mean vectors of the state output and duration distributions for training speaker $f$ and state $i$; $\boldsymbol{\mu}_i$ and $m_i$ are the corresponding means for the average voice model, and $\boldsymbol{\xi}_i = [\boldsymbol{\mu}_i^\top, 1]^\top$ and $\boldsymbol{\phi}_i = [m_i, 1]^\top$ are their correspondents in homogeneous coordinates. Finally, $\boldsymbol{W}^{(f)} = [\boldsymbol{\zeta}^{(f)}, \boldsymbol{\epsilon}^{(f)}]$ and $\boldsymbol{X}^{(f)} = [\chi^{(f)}, \nu^{(f)}]$ are transformation matrices which indicate the difference between training speaker $f$ and the average voice.

Let $F$ be the number of training speakers, $\boldsymbol{O} = \{\boldsymbol{O}^{(1)}, \ldots, \boldsymbol{O}^{(F)}\}$ be all the training data, and $\boldsymbol{O}^{(f)} = \{\boldsymbol{o}_{1_f}, \ldots, \boldsymbol{o}_{T_f}\}$ be the training data of length $T_f$ for speaker $f$. The HSMM-based speaker-adaptive training algorithm described by Yamagishi and Kobayashi (2007) simultaneously estimates the parameter set of the HSMM $\lambda$ and the set of transformation matrices $\Lambda^{(f)} = (\boldsymbol{W}^{(f)}, \boldsymbol{X}^{(f)})$ for each training speaker so that the likelihood of $\boldsymbol{O}$ is maxi-

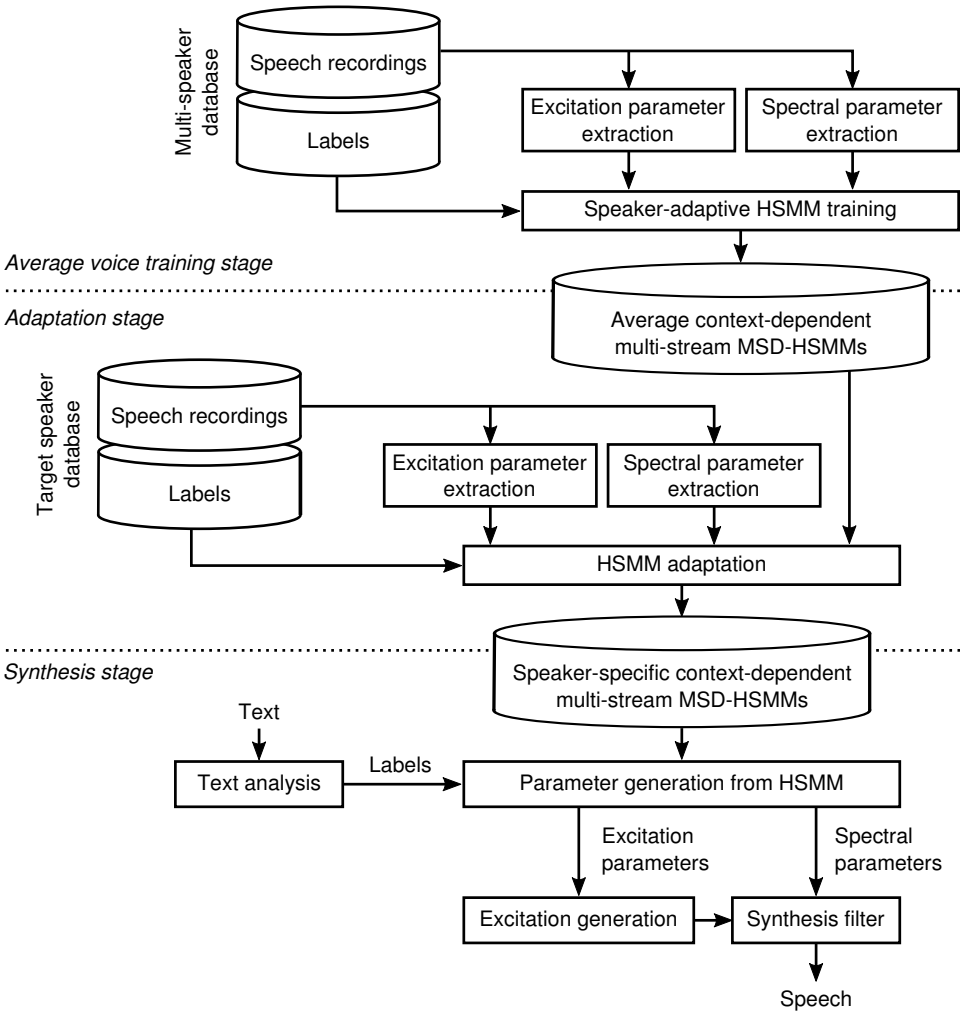



Figure 3.8: Schematic overview of a speaker-adaptive HSMM-based speech synthesis system (figure after Yamagishi et al., 2009b).

mized, i.e., it solves

$$(\lambda^*, \Lambda^*) = \underset{\lambda,\Lambda}{\operatorname{argmax}} P(\boldsymbol{O} \mid \lambda, \Lambda) \tag{3.31}$$

$$= \underset{\lambda,\Lambda}{\operatorname{argmax}} \prod_{f=1}^{F} P(\boldsymbol{O}^{(f)} \mid \lambda, \Lambda^{(f)}), \tag{3.32}$$

where $\Lambda = (\Lambda^{(1)}, \ldots, \Lambda^{(F)})$ is the set of the transformation matrices for all training speakers.

Instead of the simple MLLR method, in which only the mean vectors but not the variances are transformed, Yamagishi et al. (2009b) have proposed

feature-space speaker-adaptive training where both means and variances are affected by the transform. During clustering of the average voice models, it has been shown to be beneficial to apply node splits only if there is data from all speakers available for the resulting child nodes (Yamagishi et al., 2002; Yamagishi et al., 2003).

For adaptation of the resulting average voice model to a new target speaker, a procedure similar to SAT is applied: Using the target speaker speech data, those linear transformations that need to be applied to the average voice distributions are determined that maximize the likelihood of the adaptation data. Again MLLR may be used, but several other adaptation approaches have been investigated by Yamagishi et al. (2009a), including Constrained Maximum Likelihood Linear Regression (CMLLR), where the same transformation is applied to both mean and variance, and also a new method called Constrained Structural Maximum A Posterioi Linear Regression (CSMAPLR). CSMAPLR is shown to be particularly successful in this task, in terms of both objective and subjective evaluation results.

With as little as 50–100 sentences of adaptation data from the target speaker, natural sounding speech with high similarity to the target speaker can be synthesized using this approach (Yamagishi and Kobayashi, 2007). Interestingly, subjective listening tests have shown that an adapted voice using 100 sentences for adaptation was even rated as having more similar speaker characteristics to the target speaker than a "conventional", speaker-dependent model trained with 450 sentences from the target speaker (Yamagishi and Kobayashi, 2007). The speaker-adaptive approach was also shown to be significantly less negatively affected by the use of noisy, inconsistent and unbalanced data than other TTS paradigms (Yamagishi et al., 2009b).

In addition to adaptation towards a certain target speaker, these techniques can also be applied to other tasks such as synthesizing speech with various speaking styles (Tachibana et al., 2006), fast speech synthesis (Pucher et al., 2010a), synthesis of dialects with scarce resources (Pucher et al., 2010b; Toman et al., 2013b), and cross-dialect transformation (Toman et al., 2013a), for example.

# Chapter 4

# Developing an Audiovisual Speech Synthesis Pipeline

This chapter describes in full detail the pipeline for audiovisual speech synthesis developed and used in the research for this dissertation, from the recordings of the raw speech and motion data, via data post-processing, feature extraction, model training, and feature generation to the final animation rendering. This is not only important for understanding how such a system works and for reporting what exactly was developed within this dissertation project, but also for the reproducibility of the research. Given the information provided in this chapter, anyone should be able to repeat the experiments discussed in the following chapters and verify the results.

Parts of this chapter have been published before in Schabus et al. (2012a), Schabus et al. (2013) and Schabus et al. (2014b).

## 4.1 Equipment for Recording Speech and Motion

A marker-based optical motion capturing system called NaturalPoint OptiTrack Expression[1] was used for recording facial motion. A photograph of the system's hardware and a screen shot of its software are shown in Figure 4.1. This system consists of six FLEX:V100R2 cameras placed around the speaker's face such that different views of the same scene are captured. The cameras are equipped with infrared LEDs and they capture light in the infrared spectrum only, such that reflective materials will appear very bright in the cameras' sensors, and everything else—in particular, the speaker's face and body—will appear relatively dark. Small plastic semi-spheres with

[1] http://www.naturalpoint.com/optitrack/

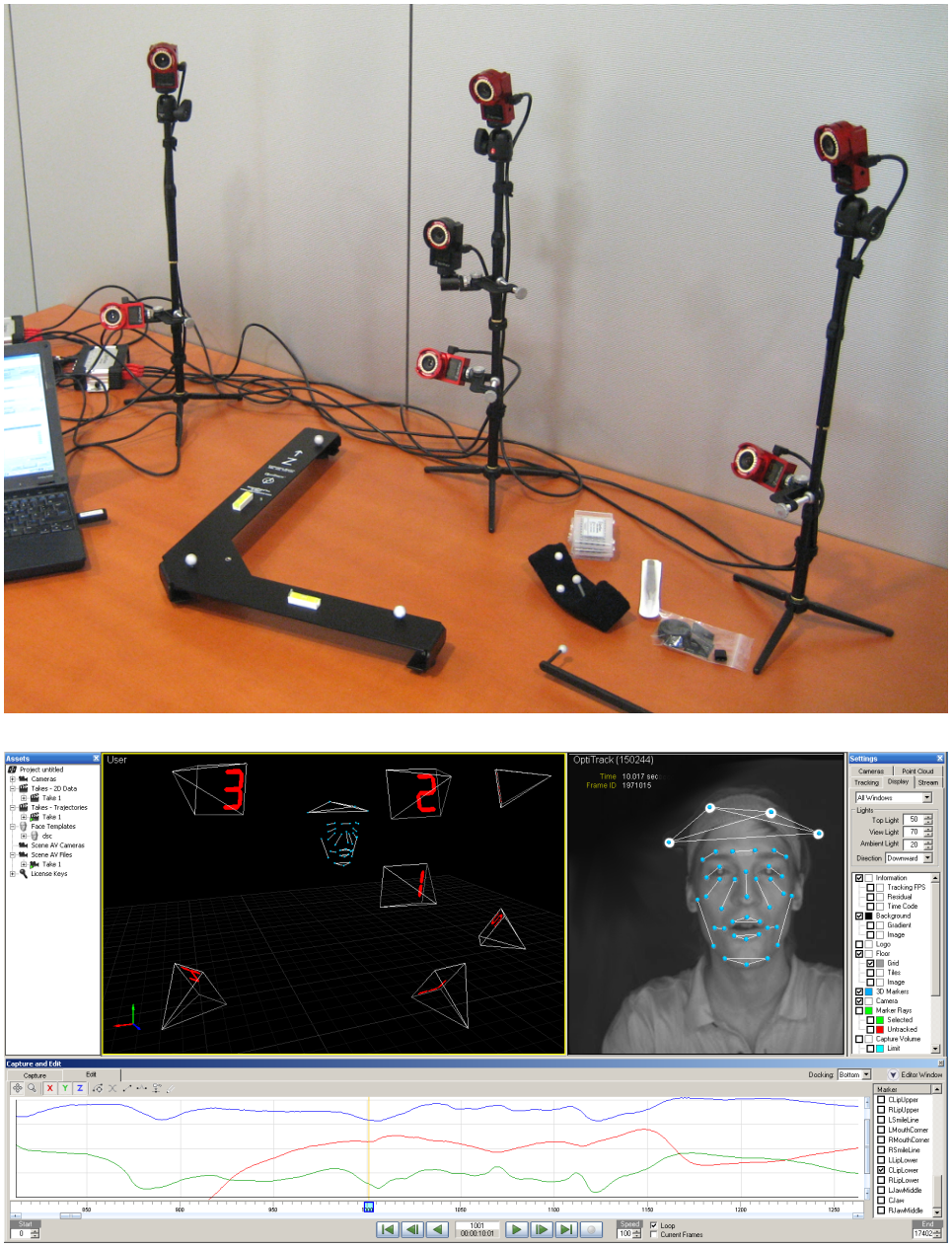



Figure 4.1: OptiTrack motion capturing hardware (top) and software (bottom).

a highly reflective surface are glued to the speaker's face, which will then appear as very bright dots in the camera images. These markers are then tracked over time in 3D space by the system in the following way: From the camera input gathered in a calibration procedure, the system reconstructs the relative position of the cameras to each other in three-dimensional space. When a marker is placed in front of the calibrated system such that it is visible from multiple cameras, the respective $x, y$ positions in these cameras' sensor planes are used together with the calibration results to compute a 3D position $x, y, z$ via triangulation. Since the time-synchronized cameras record a new image every 10 ms, the displacement of every facial marker from one frame to the next will be quite small, which allows the system to robustly track the movements of all markers over time. Thus, the system's output consists of a 3D trajectory (i.e., a sequence of 3D points) at 100 Hz for each marker.

A seventh FLEX:V100R2 camera provides 640×480 pixels grayscale video footage for control purposes, at the same frame rate of 100 Hz and with frame-level synchrony to the 3D data. Figure 4.2 shows an image from this seventh camera, also displaying the marker layout used in the recordings. 37 markers are placed on the speaker's face at the following positions:

- 5 markers down the middle of the face
    1. Nose bridge
    2. Nose tip
    3. Upper lip
    4. Lower lip
    5. Chin
- 16 markers on either side of the face
    1. Inner brow
    2. Middle brow
    3. Outer brow
    4. Upper eyelid
    5. Lower eyelid
    6. Outer eye corner
    7. Left resp. right nose bridge
    8. Cheek
    9. Ear (close to the ear on the face, not actually on the ear)
    10. Sneer line
    11. Left resp. right upper lip
    12. Left resp. right lower lip
    13. Mouth corner
    14. Smile line
    15. Inner jaw

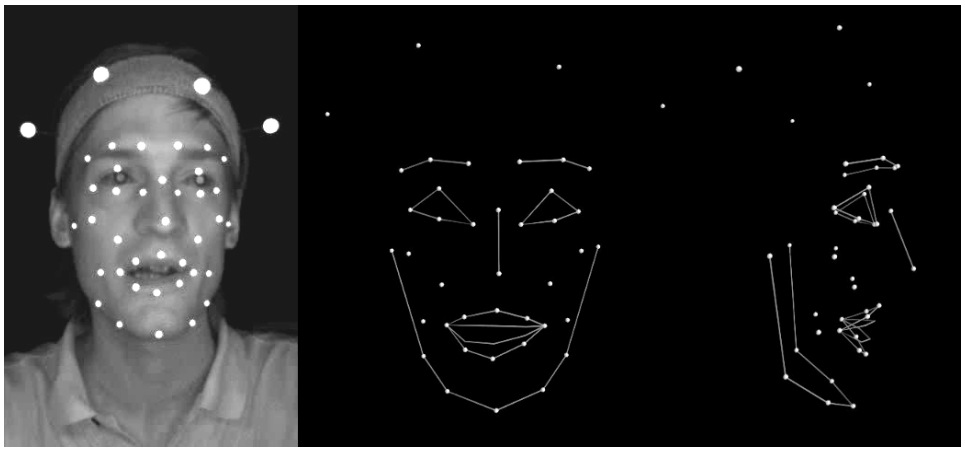


Figure 4.2: Example frame from the OptiTrack system: grayscale image from the control video (left) and reconstructed 3D points seen from a frontal view (middle) and from a side view (right). Lines added for illustrative purposes.

16. Outer jaw

A headband holds four additional markers, giving a total of 41 markers recorded by the OptiTrack system.

Although the OptiTrack system generally records the markers' trajectories quite robustly, the data does sometimes contain tracking errors. For example, when two markers come close together and subsequently depart again from each other, the system sometimes mistakes one for the other. Such "swaps" and other discontinuities like short gaps, erroneous trajectory jumps, etc. need to be manually identified and corrected. The OptiTrack software provides some tools for manually editing the recorded trajectories to that end. With a growing amount of data, manual clean-up can become very time-consuming; however, experience throughout this dissertation project has shown that starting from a "cleaner" data corpus can greatly improve the modeling and synthesis results.

For audio, a high-definition audio recorder was used (an Edirol R-4 Pro) to record signals at 44 100 Hz sampling rate, 16 bit encoding, captured by a professional microphone (an AKG C-414 B-TL). The recordings were performed in an anechoic, acoustically isolated room with artificial light only.

For synchronization between the audio and 3D recordings, the speakers were asked to produce a simple clapping signal at the beginning of each recording. This makes it straightforward to identify the position of the signal in both the audio recordings as well as in the grayscale video. Due to the frame rate of the latter, this results in a synchronization accuracy of $\pm 5$ ms. Each

recording was started in a neutral pose (relaxed face, mouth closed, eyes open, looking straight ahead).

## 4.2 Retargeting: Using Recorded Marker Motion to Control a 3D Head Model

With marker-based systems like OptiTrack, the assumption is made that the motion of the markers placed at the selected positions is sufficient to capture all facial movement and deformation. This is convenient because the markers are much easier to track robustly than natural facial feature points, and because it results in a compact and well-defined data representation. Depending on the number of markers and the chosen marker layout, as well as on the required granularity of the recorded motion, this assumption may or may not be justified. However, the use of this technique in the animation industry (Pighin and Lewis, 2006) indicates that with accurate tracking of a certain number of markers, the captured motion detail can be "good enough" for practical purposes, thus justifying the assumption to some degree.

However, even if the assumption is indeed justified, there still remains a problem to be solved, namely, how to obtain a complete animation of a high-resolution head model (i.e., a sequence of 3D coordinates for *all the vertices* of the head mesh) from recorded marker motion (i.e., a sequence of 3D coordinates for *a small number of selected vertices* of the head mesh). It is straightforward to (manually) establish a one-to-one correspondence between the recorded markers and vertices on a given high-resolution head model, even if the physiology of the recorded speaker and the 3D head are quite different. With this correspondence, the recorded marker motion already specifies the motion of some of the markers of the 3D model. However, it is much less straightforward to determine appropriate motion for the many other vertices of the head model. This process is called retargeting in animation industry parlance and Pighin and Lewis (2006) describe several techniques for retargeting a recorded facial performance (referred to as the *source* animation) onto a digital face (referred to as the *target*).

In general, the retargeting function, which maps each frame of the source into a frame for the target, needs to be crafted manually to some degree by an animation expert. This may include fine details like skin wrinkling or lip protrusion which are not actually captured by the marker motion, but which can be expected to result from a given marker constellation being applied to a given head model. It is important that the marker layout is designed such that it captures as many degrees of freedom of the recorded face as completely as possible; additional details can be inferred from physical/anatomical constraints and if they are built into the retargeting function

or the parametrization of the target head, then they can still be realized, even if they are not captured directly by the markers. Although the creation of the retargeting function involves manual input, it can be applied in a fully automatic fashion, once it is defined, to turn recorded marker data into target animations. For relatively small amounts of animation (like for an animated film of say 90 minutes), additional manual post-processing is often carried out in order to achieve supreme quality.

The typical use case for a system like OptiTrack is to record facial performances with it, which can then be used in a computer game or animated film after retargeting. This approach might reach better animation quality at reduced production costs in comparison with creating all animations entirely by hand (Pighin and Lewis, 2006). This is taken a step further by adding the concept of synthesis to the picture: instead of recording all the required performances and retargeting them, the idea is now to record marker data for a representative collection of utterances, which can be used to train a text-to-marker-data system. That system can then create appropriate marker motion data for any given text input, and together with the retargeting procedure this results in a complete text-to-visual-speech system. This can imply a drastic further reduction of the production costs per second of speech animation, especially when the number of required utterances becomes very large, and it even allows utterances to be dynamically composed during run-time. Realistically, it will however also imply a quality reduction, because it should be expected that the animations generated by a synthesis system do not reach the quality of recordings.

In Chapter 2, systems were mentioned that produce visual speech for acoustic speech input (e.g., L. Wang et al., 2010, Tamura et al., 1998a, Hofer et al., 2008 and Tao et al., 2009). These systems can be chained with a TTS system to produce speech from text, and then visual speech from the result, to provide a text-to-audiovisual speech system. However, they can also be used with recorded speech. At the moment, this may be the most realistic way to reduce costs while achieving acceptable audiovisual speech quality, because the acoustic speech signal can contain any emotional or conversational aspects required for the respective story (which is still very difficult to achieve with a TTS system), but the facial animations are created automatically. For this kind of pipeline, the step from research prototypes to commercial applications has already been taken.[2]

NaturalPoint, the manufacturer of the OptiTrack system, provides a free "Motion Builder Goody Pack"[3] which includes the 3D head model depicted

[2] For example, the company Speech Graphics shows some impressive demonstrations at http://www.speech-graphics.com/samples/.

[3] The goody pack can be downloaded from http://www.naturalpoint.com/optitrack/downloads/expression.html and some explanatory videos are available at http://www.

Figure 4.3: NaturalPoint head model showing the polygon topology, shaded without texture, shaded with texture and marker positions visible, and shaded with texture with mouth open during speech.

in Figure 4.3 consisting of 39 622 polygons, and also a ready-to-use retargeting function for applying marker motion recorded with the OptiTrack system to this head model using the Autodesk MotionBuilder[4] animation software. This head model and retargeting function were used unaltered for all experiments in this dissertation, which focuses on the task of synthesizing marker motion sequences of high quality.[5] Building a realistic head model with detailed parametrization and retargeting is considered a separate problem that is important (and also interesting), but one that is not addressed here.

## 4.3 Data Post-Processing

The recorded audio data is in standard waveform audio format and directly ready for use in the HTS system, which expects such audio files and extracts spectral parameters, fundamental frequency and band-aperiodicity features from them using STRAIGHT, as described in Section 3.1 (Audio Feature Extraction and Re-Synthesis). This section is therefore dealing mostly with the motion capturing data.

---

naturalpoint.com/optitrack/products/expression/tutorials.html.

[4] http://www.autodesk.com/products/motionbuilder/overview

[5] An example video showing a grayscale video, the raw recorded marker motion, and the retargeted marker motion to the 3D head is available at http://schabus.xyz/phd/retargeting.

### 4.3.1 Face Data Format Conversion

The OptiTrack motion capturing system stores the recorded facial motion data in its own format, but it can export to the open C3D[6] format as well as to the proprietary but widespread FBX format[7]. To ease processing, the data was converted to a more simplistic format. Each of the 41 markers has a name (e.g., "LMouthCorner") and a $(x, y, z)$ position for each recorded frame. Stacking all the coordinates vertically, in alphabetical marker name order, results in a column vector of $41 \cdot 3 = 123$ entries that describes a single frame. Then such frame column vectors are stacked horizontally, forming matrices of shape $123 \times n$, where $n$ is the number of frames (and $n/100$ is the duration of the utterance in seconds). Hence, each row of such an utterance matrix gives the trajectory of a certain coordinate of a certain marker over time.

### 4.3.2 Head Motion Removal

Since head motion influences the movement of all face markers, but our final goal is lip motion synthesis, we have to remove global head motion from the data. This can be done under the assumption of fixed distances between the four headband markers. A reference frame is chosen, and then for each other frame, a transformation matrix is computed which moves the headband markers of that frame to the same position as in the reference frame. By application of this transformation matrix to all 41 markers in that frame, we can eliminate global head motion, keeping only the facial deformation in the data. After this step, the four headband markers become static and can be removed. The position and orientation of the head in coordinate space is therefore determined by the reference frame, which will be different for each recording. In order to normalize the recordings in this respect, a global translation and rotation needs to be applied to each recording, such that the reference frames (in neutral pose) of all recordings are at the same position in the coordinate space and oriented in the same direction. For example, in the reference frame, the central upper lip marker is at the origin of the coordinate space and the line between the two ear markers is parallel to the $x$-axis.

### 4.3.3 Utterance Cutting

The data was recorded in blocks of 25 utterances. As already mentioned, each block was started in a neutral pose, followed by a clapping signal for

[6] http://www.c3d.org
[7] http://www.autodesk.com/products/fbx/overview

synchronization and then the 25 utterances were read by the speaker and recorded. This data needs to be cut into separate files per utterance with synchronized audio and motion data. For each 25 utterance block, the temporal position of the clapping signal was determined by manually identifying the frame in the 100 Hz grayscale video from the OptiTrack system where the hands first touch, and the frame in the 44 100 Hz audio signal where a certain amplitude threshold (depending on the recording level) is crossed. From these numbers, an offset in seconds between the two modalities is calculated. Using a sound detection algorithm on the audio recordings (as available, e.g., in the Audacity[8] open-source audio editor), the borders of all non-silent parts can be detected automatically. After manual cleanup, these borders can be used to cut the audio signal, and—together with the synchronization offset—also the motion data into separate files per utterance.

## 4.4 Feature Extraction

After post-processing, the facial motion for each recorded utterance is represented by a $123 \times n$ matrix, where $n$ is the number of frames at 100 Hz. We can also interpret this matrix as a 123-dimensional trajectory of length $n$, or as 41 3-dimensional trajectories (one per marker), or as 123 1-dimensional trajectories. Figure 4.4 shows 7 of the 123 trajectories of a recorded utterance of 2.22 seconds length. The figure shows the $y$ and $z$ coordinates of the central lower lip marker, the left mouth corner marker, and the nose tip marker. All plots range from 0 to 222 frames on the horizontal axis, and show the variation over time in millimeters in the space of the respective coordinate on the vertical axis. The vertical ranges are all of the same size, i.e., the scaling is identical for all plots. Several observations can be made from these plots:

- The trajectories vary around a certain value on the respective axis, depending on the marker's position in coordinate space (and on the face).
- The degree of variation can differ substantially between markers. E.g., in this utterance, the $z$ coordinate of the left mouth corner varies by more than 13 mm; the $y$ coordinate of the nose tip by less than 1 mm.
- The three coordinates of a given marker are generally not independent from each other. E.g., the $y$ and $z$ coordinates of the central lower lip marker show some obvious similarity.

[8] http://audacity.sourceforge.net

- The same is true also across markers. E.g., the $z$ coordinates of the central lower lip and the left mouth corner are also quite similar.

Especially the last two points are relevant for trajectory modeling with HMMs as described in Chapter 3. If considerable correlations between some of the 123 coordinates are to be expected, estimating Gaussian distributions for this data requires estimating $123 \times 123$ covariance matrices, i.e., more than 15 thousand parameters. If, on the other hand, a de-correlated representation of this data is found, with no (or at least negligible) correlation between any two coordinates, then all off-diagonal values of the covariance matrices are (almost) zero and it is sufficient to estimate the values on the diagonal, i.e., the variances of the 123 coordinates.

Furthermore, it is intuitively clear that 123 degrees of freedom are more than expected to be necessary to parametrize the possible movements of the face while speaking; there are many strong constraints on the deformation of a person's face while speaking, for example simple mechanical limitations. Therefore, a procedure to reduce dimensionality while keeping the relevant information is needed. The standard method of PCA (Pearson, 1901; Shlens, 2014) is a natural choice here, as it provides both flexible dimensionality reduction and component de-correlation. The following subsections will therefore describe how PCA can be applied to the facial motion data (as published in Schabus et al., 2012a) and justify its use for visual speech synthesis, with a focus on allowing speaker-adaptive modeling (as published in Schabus et al., 2013).

### 4.4.1 PCA-based Feature Extraction

In a single-speaker setup with $m$ utterances, the available facial motion data consists of a collection of $m$ matrices, whose rows are marker coordinate trajectories. Before proceeding to PCA, the four headband markers can be removed from the data, because they are static anyway after global head motion removal (see Section 4.3.2). Furthermore, it can be argued that eye blinking is not directly related to speech production and therefore the four eyelid markers should also be excluded. Thus 33 markers remain, and the $m$ matrices all have 99 rows (3 spatial coordinates per marker), and a varying number of columns, depending on the length of the utterance. Let us denote the matrices containing the facial motion data by $\boldsymbol{F}_i$ and their lengths by $n_i$:

$$\boldsymbol{F}_i \in \mathbb{R}^{99 \times n_i} \quad \text{for} \quad i \in \{1, ..., m\}. \tag{4.1}$$

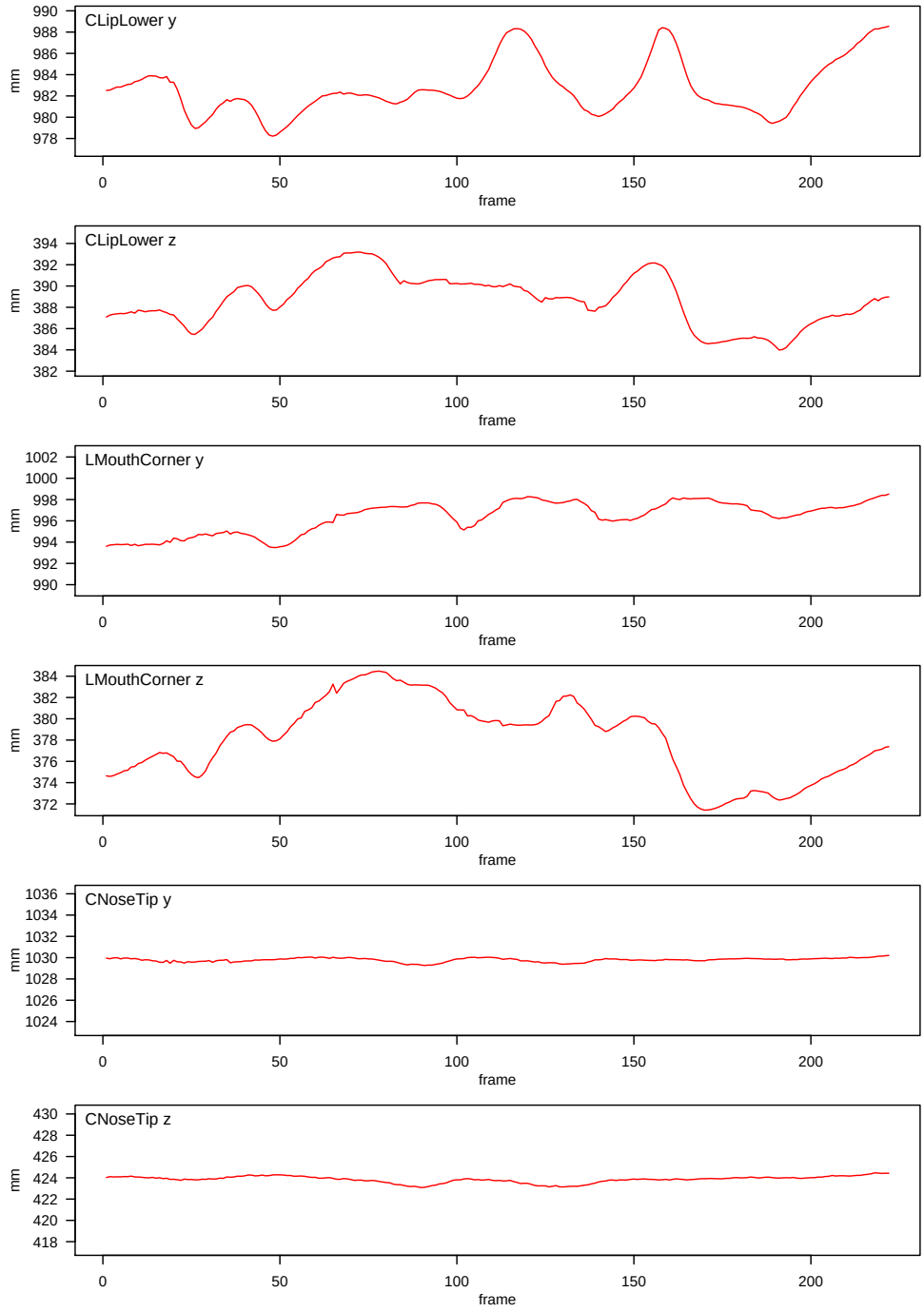



Figure 4.4: Example facial marker trajectories.

By stacking all utterance matrices horizontally, we obtain a (very wide) single matrix $\boldsymbol{A}$ of all utterances from this speaker:

$$\boldsymbol{A} = [\boldsymbol{F}_1 \boldsymbol{F}_2 \cdots \boldsymbol{F}_m] \in \mathbb{R}^{99 \times N}, \text{ with} \tag{4.2}$$

$$N = \sum_{i=1}^{m} n_i. \tag{4.3}$$

Next, we calculate the mean column vector $\boldsymbol{\mu}$ of $\boldsymbol{A}$ and subtract it from each column of $\boldsymbol{A}$ to obtain the "mean-normalized" $\bar{\boldsymbol{A}}$:

$$\boldsymbol{\mu} = \tfrac{1}{N} \cdot \boldsymbol{A} \cdot \mathbf{1}, \tag{4.4}$$

$$\bar{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{\mu} \cdot \mathbf{1}^\top, \tag{4.5}$$

where $\mathbf{1}$ denotes a column vector of $N$ ones and $\top$ denotes matrix transpose. Finally, the matrix $\bar{\boldsymbol{A}}$ can be decomposed using Singular Value Decomposition (SVD) (Shlens, 2014):

$$\bar{\boldsymbol{A}} = \boldsymbol{U} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{V}^\top. \tag{4.6}$$

We are solely interested in the matrix $\boldsymbol{U}$ of size $99 \times 99$, whose columns are the bases of the principal component space, sorted by decreasing eigenvalues. Using $\boldsymbol{U}$, we can project a frame column vector $\boldsymbol{x}$ into PCA space by multiplying $\boldsymbol{U}^\top$ from the left ($\boldsymbol{U}^\top \cdot \boldsymbol{x}$), and back into the original space by multiplying $\boldsymbol{U}^\top$'s inverse from the left. Since $\boldsymbol{U}$ is orthogonal, we have $(\boldsymbol{U}^\top)^{-1} = (\boldsymbol{U}^\top)^\top = \boldsymbol{U}$ and thus

$$\boldsymbol{x} = \boldsymbol{U} \cdot (\boldsymbol{U}^\top \cdot \boldsymbol{x}). \tag{4.7}$$

Because the bases in $\boldsymbol{U}$ are sorted, we can approximate the equality of Equation 4.7 by using only the "most relevant" components, i.e., the first $k$ columns of $\boldsymbol{U}$:

$$\boldsymbol{x} \approx \boldsymbol{U}_k \cdot (\boldsymbol{U}_k^\top \cdot \boldsymbol{x}), \tag{4.8}$$

where the quality of the approximation improves with increasing value of $k$.

Note that instead of using SVD to find the principal components, equivalently they can be found by computing the eigenvectors of the covariance matrix of the data, and that thus the PCA projection of the data amounts to a diagonalization of the covariance matrix (Shlens, 2014). Although SVD is more efficient to compute in practice, the latter view makes it clearer how PCA results in de-correlation.

So we can carry out SVD on the mean-subtracted data $\bar{\boldsymbol{A}}$ of a speaker, choose a value for $k < 99$ and project the data into a smaller ($k$-dimensional) subspace using $\boldsymbol{U}_k^\top$. Then, HMM training and synthesis can be performed

using this more compact and de-correlated representation of the speaker's data. Synthesized utterances can be projected back into the full 99-dimensional space using $\boldsymbol{U}_k$, and by re-adding the sample mean $\boldsymbol{\mu}$ we finally obtain the corresponding synthesized facial marker movement.

The subtraction of the sample mean $\boldsymbol{\mu}$ before SVD is a standard procedure before a change of basis (like PCA), but it plays an interesting role in our facial marker setup: $\boldsymbol{\mu}$ contains the average position of each marker, and subtracting it from $\boldsymbol{A}$ to obtain $\bar{\boldsymbol{A}}$ means that the latter will contain markers moving about the origin of the 3D coordinate space. In other words, $\boldsymbol{\mu}$ contains the general position of the markers, i.e., the facial geometry of the speaker, and $\bar{\boldsymbol{A}}$ contains only the motion data relative to these positions.

Furthermore, it is worth noting that the "projection function" $\boldsymbol{U}_k^\top$ can of course be applied not only to column vectors from $\bar{\boldsymbol{A}}$ (from which $\boldsymbol{U}$ is derived using SVD), but to any 99-dimensional vector, to project it into PCA space. This is of great interest for multi-speaker scenarios, in particular for speaker-adaptive modeling, where the aim is to train an average model across multiple speakers, which can later be adapted towards a previously unknown target speaker. In such a scenario, we can combine mean-normalized matrices $\bar{\boldsymbol{A}}_1, \bar{\boldsymbol{A}}_2, ...$ from multiple speakers into one big matrix $\bar{\boldsymbol{A}}_{avg}$ to carry out SVD to find $\boldsymbol{U}_k^\top$. The key idea regarding the speaker-adaptive scenario is now to apply this same projection, which was determined on the data for the average voice, to project the adaptation data from the target speaker into the same subspace. This assumes that we find a subspace via SVD on the data from the (potentially large number of) speakers in the average voice that is general enough to also contain the target speaker's data, provided that we do not choose the value of $k$ too "tight". The purpose of the following subsection is to justify this assumption, as well as to choose an appropriate value for $k$.

### 4.4.2 Objective and Subjective Feature Evaluation

The study described here has been published in (Schabus et al., 2013) and it is based on the data of three standard Austrian German speakers. In order to evaluate how well the results of PCA, when carried out the way we have described in the previous subsection, do match our task of visual feature extraction, we use both objective and subjective performance measures.

One of the three is always considered as the target speaker, i.e., the data to be projected (and reconstructed, when we consider the reconstruction error) are all frames of all utterances of that speaker. The data used for SVD, i.e., for calculating the projection function into principal component

space is either

**Method 1** the data from the target speaker

**Method 2** the data from all three speakers (including the target speaker)

**Method 3** the data from the two other speakers (excluding the target speaker).

Especially the third case is of high relevance in an adaptation scenario, as the data of the target speaker is typically not part of the training data for the average voice. Intuitively, we expect this to be the most challenging of the three scenarios. But also the second case can be of practical relevance: when we want to put all available data to optimal use, it might be beneficial to include the target speaker in the average voice.

First, we consider the objective reconstruction error. This should bring insight to the behavior of the three methods mentioned above, to understanding the role of certain markers, as well as to the influence of $k$, the number of kept dimensions. Then, the results of a subjective evaluation are presented which was carried out with 40 test subjects. The main purpose of this is to provide a basis for deciding on the value of $k$.

### Objective Evaluation via Reconstruction Error

Given a matrix $\boldsymbol{U}_k$ containing the first $k$ columns of a matrix $\boldsymbol{U}$ resulting from SVD (as described in Section 4.4.1), we define the reconstruction of a data matrix $\boldsymbol{A}$, containing a target speaker's utterances stacked horizontally, as

$$\bar{\boldsymbol{A}}_{rec} = \boldsymbol{U}_k \cdot \boldsymbol{U}_k^T \cdot \bar{\boldsymbol{A}}. \tag{4.9}$$

Re-adding $\boldsymbol{A}$'s sample mean to $\bar{\boldsymbol{A}}_{rec}$ gives us $\boldsymbol{A}_{rec}$, and we can compute the error matrix $\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{A}_{rec}$. Let $N$ denote the total number of frames in all utterances of the target speaker, i.e., $\boldsymbol{A}$, $\bar{\boldsymbol{A}}$, $\bar{\boldsymbol{A}}_{rec}$, $\boldsymbol{A}_{rec}$ and $\boldsymbol{E}$ are all of size $99 \times N$, while $\boldsymbol{U}_k$ is of size $99 \times k$. Finally, we define the reconstruction error as the Root Mean Squared Error (RMSE) across all elements $e_{ij}$ of $\boldsymbol{E}$:

$$\text{RMSE} = \sqrt{\frac{1}{99N} \sum_{i=1}^{99} \sum_{j=1}^{N} {e_{ij}}^2}. \tag{4.10}$$

We have computed the RMSE for all $k \in \{1, \ldots, 99\}$ and for each of the nine conditions resulting from the combination of each of our three speakers as target speaker with one of the three methods to compute the SVD as described above. The results are shown in Figure 4.5. The points are labeled

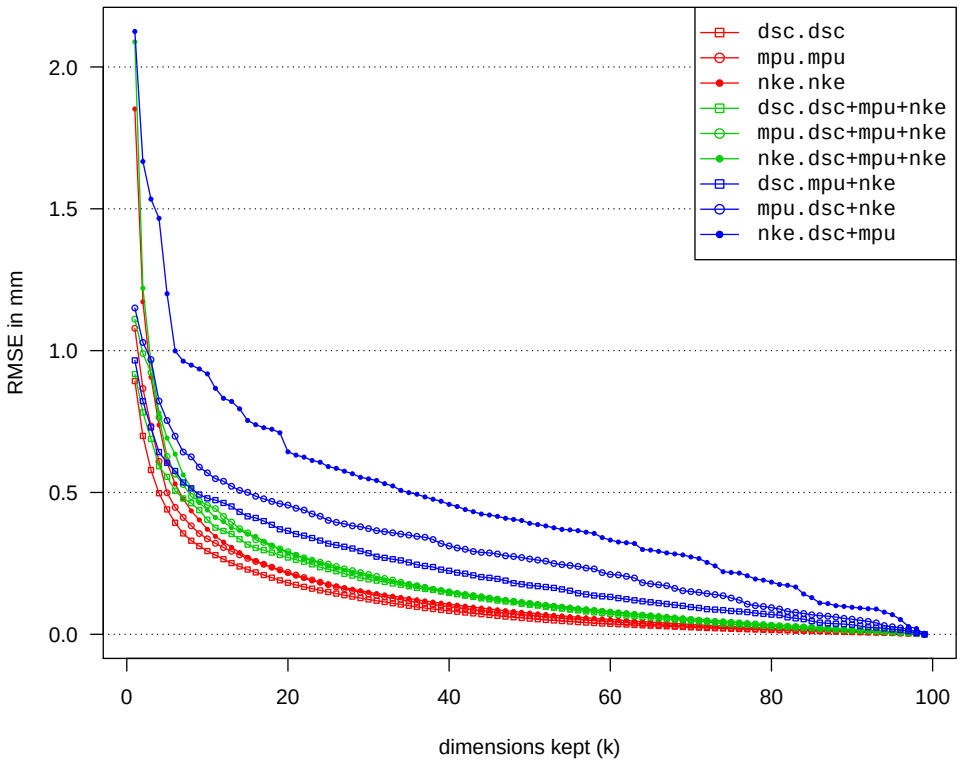



Figure 4.5: PCA reconstruction error (RMSE) for the nine different conditions and varying $k$. In the labels of the legend, the identifier of the target speaker is given before the period, and the identifiers of the speakers used to compute $\boldsymbol{U}$ via SVD are given after the period.

with the target speaker before the period and all speakers that were used in the SVD after the period.

Overall, we see our intuition confirmed: using only 6 of 99 dimensions yields an RMSE of less than 1 mm in all nine conditions. The three speaker-specific versions (red) produce the best results, as expected. Their RMSEs lie even below 0.5 mm at $k = 6$. The three versions with all speakers in the SVD (green) are a bit worse than that, and as expected the three held-out versions (blue) yield the worst results. It takes 35 dimensions for the particularly bad *nke.dsc+mpu* to reach an RMSE below 0.5 mm.

Although the methods of the third kind produce a larger reconstruction error than the others, they still show the same overall behavior (shape of the curves in Figure 4.5), namely that the first few dimensions make a very big difference in the results, and that the error levels off towards the larger values of $k$. This means that we have the positive result that it is possible to project some speaker's data into a much smaller subspace, where the definition of

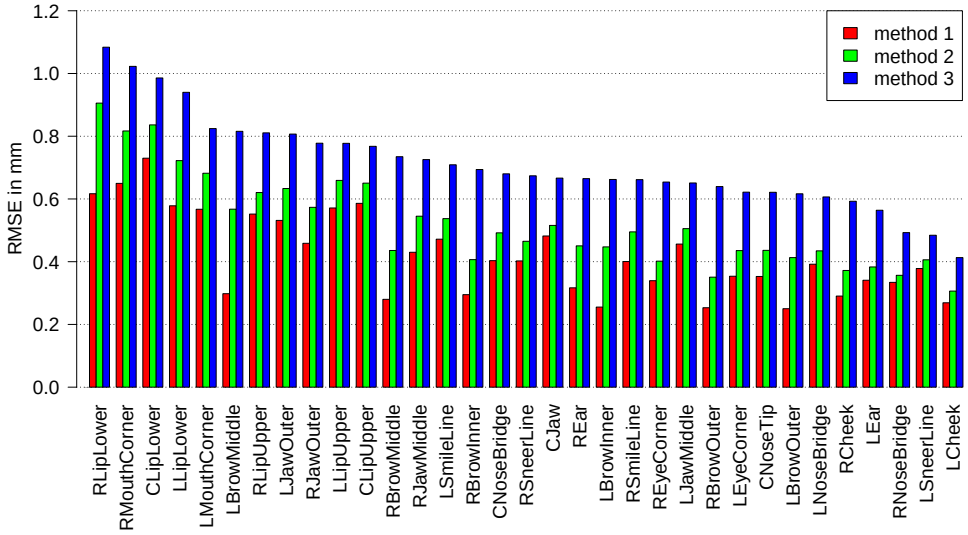



Figure 4.6: PCA reconstruction error (RMSE) for each marker and SVD method, averaged across all target speakers and 3D coordinates, at $k = 6$.

the subspace and the projection into it were determined without using any data from that speaker, without making a large reconstruction error, given that we do not choose the value of $k$ too aggressively.

Rather than taking the mean across the entire error matrix $\boldsymbol{E}$, we can also look at the means of each row, which corresponds to the mean error for a certain coordinate of a certain marker. Figure 4.6 shows the RMSE for each marker and each of the three methods (using just the target speaker for SVD, using all three speakers for SVD, using the respective other two speakers for SVD). The plotted values are means across all target speakers, 3D coordinates and of course frames, for a fixed value of $k = 6$. We can see that the markers in the region of the mouth (**Lip**, **Mouth**, **Jaw**) are responsible for the largest errors. Also, we see again how the third method (held-out) is consistently worse than the second (all speakers), which is in turn consistently worse than the first (speaker-specific).

### Subjective Evaluation via Perceptive Experiments

Based on the objective evaluation alone, it would be difficult to choose a value for $k$ to proceed to actual training and synthesis. It is not clear a priori what an RMSE of, e.g., 1 mm means perceptually, or in other words it is not clear how small we can choose $k$ without perceived degradation in quality. To clarify this, we have carried out a subjective perceptual experiment with 40 non-expert test subjects (half females and half males, aged 20–68 years). This experiment was designed as follows.

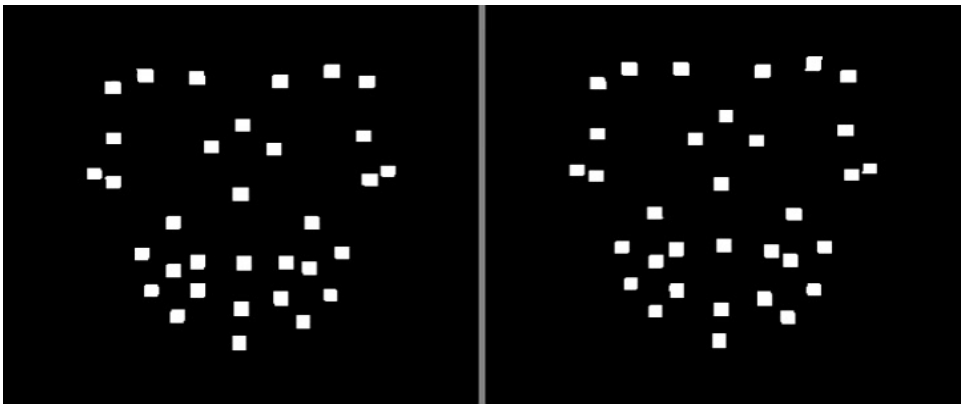


Figure 4.7: Example frame from a subjective evaluation video, showing original motion (right) and reconstructed motion (left) side by side.

We have created videos of marker renderings, where for each frame of an utterance, a white cube is drawn on a black background for each of the 33 markers at the 3D position of that marker in that frame. Note that we deliberately chose not to apply the marker motion to a virtual head and use renderings of the animated head in the evaluation, because we wanted to make sure the quality (or lack of quality) of the retargeting or the visual appearance of the head do not skew the evaluation results. In each video, we showed a rendering of the originally recorded data side by side with a rendering of a reconstruction using a certain value of $k$. This leads to renderings as the one shown in Figure 4.7.[9] Then the test subjects were asked to decide whether the two renderings were *different* or *the same* from their point of view. Whether the original was on the left or on the right was chosen randomly for each video.

We used the first five sentences of our corpus as test sentences, and each test subject saw one comparison for each test sentence and each of the nine conditions (cf. Figure 4.5), i.e., 45 comparisons in total. We have selected values of $k$ with respect to the reconstruction error: For each of the nine conditions, we have partitioned the set of 99 possible values for $k$ into 19 bins, where each bin amounts for a similar percentage of the overall error. We also added a twentieth bin containing only the last value ($k = 99$). We then selected the middle value of each bin as its representative. Each test subject saw at least one comparison from each bin, with the remaining comparisons distributed randomly. Table 4.1 shows for target speaker *nke* which values of $k$ belonged to which of the 20 bins in each of the three methods.

This leads to a denser sampling in the lower region, where one additional

[9] A video showing examples is available at http://schabus.xyz/phd/pcaevaluation.

Table 4.1: Partitioning of the values of $k$ for target speaker *nke*.

| Bin | Method 1 | Method 2 | Method 3 |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2–3 |
| 3 | 3 | 3 | 4–5 |
| 4 | 4 | 4–5 | 6–8 |
| 5 | 5–6 | 6–7 | 9–11 |
| 6 | 7–8 | 8–9 | 12–14 |
| 7 | 9–10 | 10–11 | 15–17 |
| 8 | 11–12 | 12–14 | 18–21 |
| 9 | 13–15 | 15–17 | 22–25 |
| 10 | 16–18 | 18–20 | 26–29 |
| 11 | 19–21 | 21–24 | 30–33 |
| 12 | 22–25 | 25–28 | 34–38 |
| 13 | 26–29 | 29–33 | 39–43 |
| 14 | 30–34 | 34–38 | 44–49 |
| 15 | 35–40 | 39–44 | 50–55 |
| 16 | 41–47 | 45–52 | 56–62 |
| 17 | 48–56 | 53–62 | 63–70 |
| 18 | 57–70 | 63–78 | 71–80 |
| 19 | 71–99 | 79–99 | 81–99 |
| 20 | 99 | 99 | 99 |

dimension makes a big difference, and a sparser sampling in the higher region, where one additional dimension makes a small difference. The entire evaluation thus amounts to 900 comparisons (9 methods $\times$ 5 sentences $\times$ 20 bins), for each of which we have two votes from two different subjects. Therefore the results contain a total of 1800 votes.

The results are shown in Figure 4.8, where we have plotted the percentage of "different" votes for each of the 20 bins (top), and the same data additionally separated by method (bottom). We see that the general picture is in agreement with what we know from the objective error, namely that reconstructions with low values of $k$ (left side of Figure 4.8) are perceived as being mostly different from the original, that small changes to $k$ have a shrinking influence with growing $k$, and that the difference levels off towards the upper end of the scale (right side of Figure 4.8).

However, the actual values of the evaluation at the extreme points are somewhat surprising: The reconstructions corresponding to the first bin are very poor in terms of the objective error and should look clearly different from the original, yet in 13 of the 90 comparisons (14 %) they were perceived as being equal by the test subjects. Similarly, at the other end of the scale, the reconstructions with $k = 99$ in bin 20 are per definition error-free, as the projection into principal component space and back are mere rotations

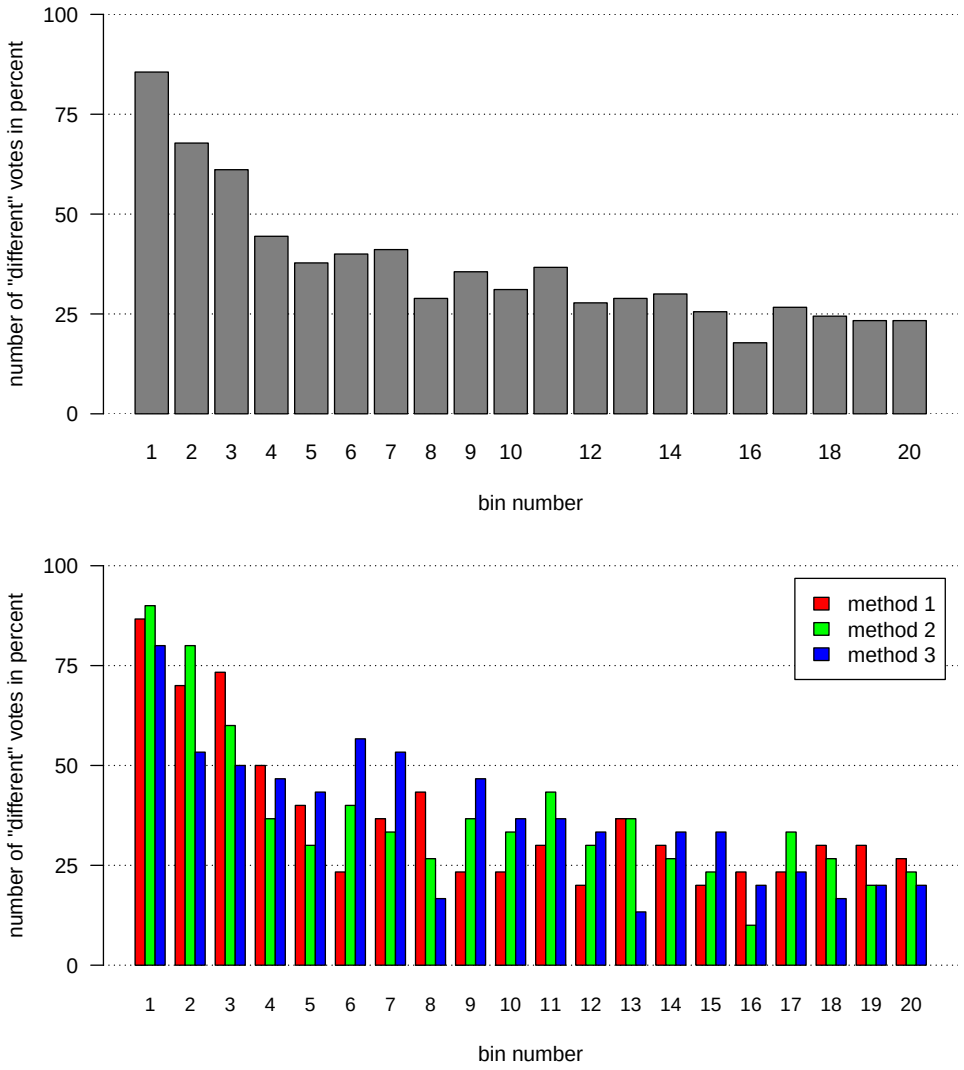



Figure 4.8: Results of the subjective evaluation: Percentage of "different" votes per bin (top) and per bin and method (bottom).

of the coordinate system.[10] Nevertheless, in 21 out of 90 cases (23 %) they were judged as being different from the original.

We believe some of this uncertainty in the results can be ascribed to the difficulty of the task. Even if the marker motion is quite different from the original for low values of $k$, the overall appearance of the two renderings is very similar. Furthermore, the sequence of comparison examples is quite uniform, which could lead to effects of boredom.

This uncertainty also makes it difficult to compare the three methods to

[10] The actual RMSE in our implementation was always $< 10^{-15}$.

Table 4.2: Significant differences in perception: Results of paired Wilcoxon signed rank tests between votes for each bin, with (■) and without Bonferroni correction (□). The symbol "·" indicates no significant difference.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | · | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 2 | □ | · | · | □ | ■ | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 3 | ■ | · | · | □ | □ | □ | □ | ■ | □ | ■ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 4 | ■ | □ | □ | · | · | · | · | □ | · | · | · | □ | □ | □ | □ | ■ | □ | □ | □ | □ |
| 5 | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | □ | · | □ | □ | □ |
| 6 | ■ | □ | □ | · | · | · | · | · | · | · | · | · | · | · | □ | □ | · | □ | □ | □ |
| 7 | ■ | □ | □ | · | · | · | · | · | · | · | · | · | · | · | □ | □ | □ | □ | □ | □ |
| 8 | ■ | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · |
| 9 | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | □ | · | · | · | · |
| 10 | ■ | ■ | ■ | · | · | · | · | · | · | · | · | · | · | · | · | □ | · | · | · | · |
| 11 | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | □ | · | □ | □ | □ |
| 12 | ■ | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · |
| 13 | ■ | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · |
| 14 | ■ | ■ | ■ | □ | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · | · |
| 15 | ■ | ■ | ■ | □ | · | □ | □ | · | · | · | · | · | · | · | · | · | · | · | · | · |
| 16 | ■ | ■ | ■ | ■ | □ | □ | □ | · | □ | □ | □ | · | · | · | · | · | · | · | · | · |
| 17 | ■ | ■ | ■ | □ | · | · | □ | · | · | · | · | · | · | · | · | · | · | · | · | · |
| 18 | ■ | ■ | ■ | □ | □ | □ | □ | · | · | · | □ | · | · | · | · | · | · | · | · | · |
| 19 | ■ | ■ | ■ | □ | □ | □ | □ | · | · | · | □ | · | · | · | · | · | · | · | · | · |
| 20 | ■ | ■ | ■ | □ | □ | □ | □ | · | · | · | □ | · | · | · | · | · | · | · | · | · |

each other based on the subjective data. The bottom part of Figure 4.8 illustrates that the data does not allow for drawing clear conclusions in this regard.

To assess the statistical significance of the differences between the bins' results, we have computed Bonferroni-corrected paired Wilcoxon signed rank tests between the votes of each pair of bins. The pairing of votes was based on the method and utterance only, i.e., we ignored which test subject cast a particular vote. The results are shown in Table 4.2, where the symbol "■" indicates a significant difference ($\alpha = 0.05$). In this rather restrictive setting (due to Bonferroni correction the value of $\alpha$ for each of the 190 tests is $0.05/190 \approx 0.00026$), only the first four bins show significant differences from some of the other bins, i.e., none of the bins from 5 to 20 differ significantly from each other.

This result tells us that we need to choose $k$ from a bin $\geq 4$ at the very least, and it even suggests that choosing from bin number 4 is sufficient, since larger values do not lead to significantly better results anyway. However, the conservativeness of Bonferroni correction would act in our advantage here, because it reduces the probability of false positives (type I error) at the cost of an increased probability of false negatives (type II error). We should

not choose $k$ too small because of some significant differences that were missed due to the Bonferroni correction. Therefore, Table 4.2 also shows the additional significances of the same test without Bonferroni correction, indicated by the symbol "□". This result is quite likely to contain some false positives, but there is nevertheless the set of bins $\{12, ..., 20\}$ where there are no significant differences. Therefore, by selecting the smallest $k$ larger than any $k$ from bin 11 ($k = 33$) we still make a conservative choice. However, the final $k = 33$ still accounts for a great reduction in dimensionality: Two thirds of the initial 99 degrees of freedom could be removed.

#### Discussion

Overall, both the objective and the subjective evaluation have provided results in general agreement with the expectations. With growing $k$, the results improve quickly at first, and finally level off—towards zero in the objective case and towards "background noise" of uncertainty in the subjective case. The reconstruction error evaluation clearly showed the difference in performance between the three methods, something which the subjective method failed to show. However, the user votes provide an excellent basis for selecting an actual value for $k$ that defines the number of dimensions employed in both training and synthesis.

## 4.5 Model Training and Synthesis of Speech Motion

After we have defined the visual features as discussed in the previous section, we can view the steps of computing the PCA on the marker data and the projection into a reduced $k$-dimensional PCA space as the feature extraction procedure for the visual data, similar to the speech feature extraction discussed in Chapter 3: Some computation is applied to the recorded signal, resulting in feature vector sequences; these are used as training data to estimate HSMM parameters; synthesizing from the HSMMs will again produce such feature vector sequences for new textual input; and finally the inverse of the analysis procedure is applied to turn the synthesized feature vectors into a full signal again.

Because the visual recordings are already of relatively low temporal resolution (100 Hz) there is no need for further reduction. As discussed in Chapter 3, the feature extraction procedure for the acoustic speech signal resulted in a change from 44 100 Hz in one dimension to 200 Hz in 66 dimensions. In order to match the temporal resolution of the audio features, the

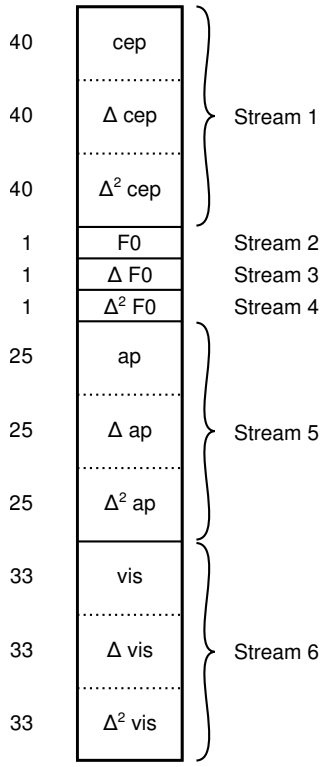



Figure 4.9: Structure of an audiovisual feature vector consisting of acoustic speech features and visual speech features of one point in time.

visual features can be "inflated" from 100 Hz to 200 Hz using cubic interpolation. After this, a combined observation vector consisting of spectral, fundamental frequency, aperiodicity and visual features can be assembled for each point in time, as depicted in Figure 4.9 (cf. Figure 3.4). Just like for the acoustic features, first and second order time differences are added for the visual features to capture the dynamics of the signal.

Using such observations, it is fairly straightforward to train a text-to-audiovisual speech synthesizer. The HSMM training procedure needs to be modified to include this additional feature stream, with separate clustering trees for the visual features (just like the different acoustic features, which are also clustered independently). We may also simply build an observation consisting of visual features only, for a text-to-visual speech system. In either case, after training we may carry out text-to-observation sequence synthesis in the same way as for audio-only modeling, as described in Chapter 3. The generated PCA-space vectors can be re-projected into the full-dimensional space using the projection matrix from SVD before training, resulting in facial marker trajectories. These can be applied to a 3D head model using a retargeting function, to obtain a 3D facial animation, which can finally be

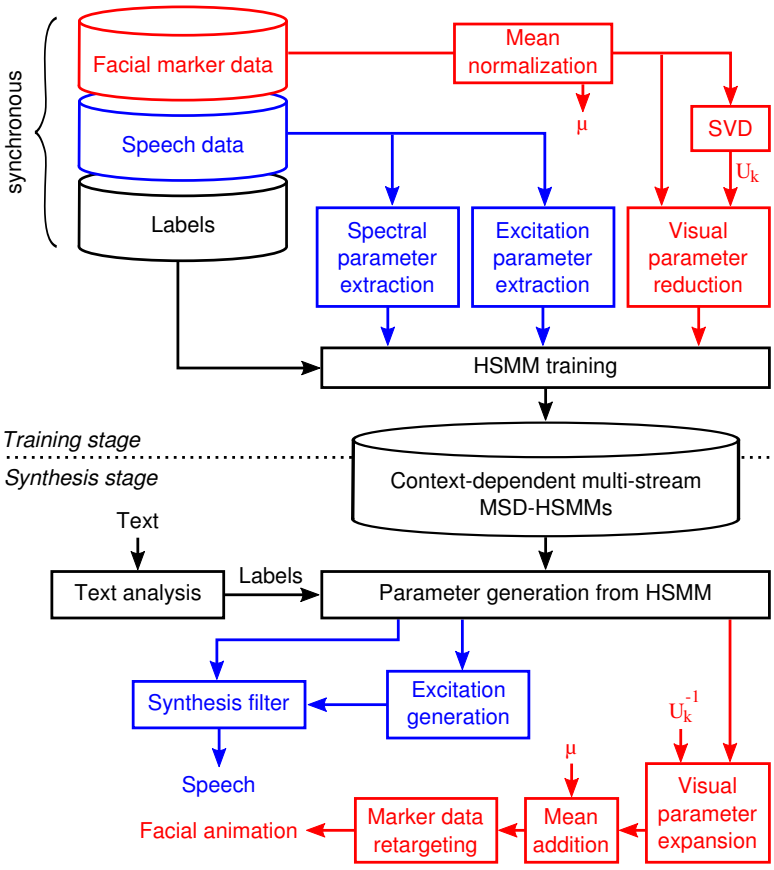



Figure 4.10: Overview of a speaker-dependent audiovisual speech synthesis system, which consists of three main components: audiovisual speech analysis, audiovisual training, and audiovisual speech generation. The corresponding audio-only system does not include the red parts, and the corresponding visual-only system does not include the blue parts.

rendered to a video file or plugged into a 3D game, etc. Figure 4.10 provides an overview of the audiovisual HSMM-based system. By removing the red parts of the figure, an audio-only system is obtained (identical to Figure 3.7 from Chapter 3). By removing the blue parts of the figure, a visual-only system is obtained.

For the experiments in this dissertation, the training scripts published by the Centre for Speech Technology Research (CSTR) of the University of Edinburgh in the project Effective Multilingual Interaction in Mobile Environments (EMIME)[11] have been adapted for visual/audiovisual training and synthesis. The CSTR/EMIME scripts use HTS version 2.1.

[11] http://emime.org

Before concluding this chapter on the audiovisual speech synthesis pipeline, it should be emphasized once more that the investigations in this dissertation focus on speech motion trajectory modeling and synthesis, and that the construction of a detailed 3D face representation as well as the problem of detailed deformation driven by low-dimensional control parameters are considered separate problems which are not addressed in this dissertation. In particular, the trajectories used for training and synthesis come from a marker-based facial motion capturing system, and the problem of retargeting such trajectories to a high-resolution head model is solved by a given retargeting function included with the head model distributed by the tracking system manufacturer. However, the applied HSMM training and synthesis procedures are not limited to this particular setup, the same methodology can be used for other time series representations of speech motion, coming from other sources. For example, given a collection of speech animation in the form of blend shape parameter sequences (which are common in the animation industry), either hand-animated or based on some performance capturing technology, could be used to train an HSMM-based synthesis system in the exact same way.

# Chapter 5

# Audiovisual Speech Corpora

This chapter describes the audiovisual speech data corpora that were created for the research in this dissertation: 1) One corpus of three speakers from the project team, each reading 223 Standard Austrian German utterances, 2) one corpus of eight speakers from two specific Austrian towns, each producing around 650 utterances in their respective dialect plus 223 utterances in Standard Austrian German, and 3) a corpus of a single speaker reading 320 Standard Austrian German utterances at normal, fast and slow speaking rate. All three were recorded as described in the previous chapter, i.e., using the OptiTrack system for recording facial motion with synchronous studio-quality audio recordings. For the third corpus, an electromagnetic articulator tracking system was additionally used to record tongue motion. Parts of this chapter have been previously published in Schabus et al. (2012a), Schabus et al. (2014a) and Schabus et al. (2014b).

For data-driven speech technology research, training corpora of speech data are an essential asset that is often created and used by research groups when required, but less often made available for the general research community. The creation of high-quality annotated corpora is a highly time-consuming and hence expensive task. This is true to an even larger extent when multiple modalities are recorded simultaneously, because of the additional requirement of synchronization between the different modalities. Furthermore, corpora including data acquired using special hardware, like motion capturing and Electro-Magnetic Articulography (EMA), are even more expensive to create because the equipment itself and the know-how to operate it are required for recording. To our knowledge, there are only two corpora of EMA data available free of charge, both from the University of Edinburgh (Wrench, 1999; Richmond et al., 2011). As far as speech fa-

cial motion capture data is concerned, there is for example a corpus of 10 speakers in affective dyadic interaction in American English (Busso et al., 2008).[1]

In order to improve this situation, the first and third corpus have already been made publicly available on the Internet, free of charge, for research purposes. The dialect speaker corpus will also be released in a similar fashion at a later point in time.

## 5.1 Face Motion and Speech Corpus (FMSC)

In a first recording round, three members of the project team (one female and two males, including the author) were recorded, reading the same recording script in standard Austrian German. The script contains sentences from a well-known German text corpus (100 "Berlin" sentences, 100 "Marburg" sentences, 16 "Buttergeschichte" sentences, 7 "Nordwind und Sonne" sentences). It is phonetically balanced, i.e., it contains all phonemes in relation to their appearance in German, and it contains utterances of varying length, to cover some prosodic variance (phrase breaks, etc.). It amounts to 223 utterances and roughly 11 minutes total for each of the speakers. Speech and facial motion were recorded as described at the beginning of Chapter 4 (facial motion capturing with OptiTrack system, studio-quality audio). A description of the FMSC corpus was published by Schabus et al. (2012a), and the data is available from `http://schabus.xyz/phd/fmsc`.

## 5.2 Bad Goisern and Innervillgraten Dialect Speech Corpus (GIDS)

To investigate the influence of dialectal variation on audiovisual speech, eight dialect speakers were recorded in the AVDS project: Two female and two male speakers from Bad Goisern, a town in the Salzkammergut region; and two female and two male speakers from Innervillgraten, a town in the Osttirol region. The variety of German spoken in Bad Goisern belongs to the Central Bavarian group, whereas the variety spoken in Innervillgraten belongs to the Southern Bavarian group (Hornung et al., 2000), as illustrated in Figure 5.1. The two dialects are quite different from each other and both are also quite different from standard Austrian German. A recording script for each of the two variants was created within the project AVDS, amounting to 665 utterances for the Bad Goisern dialect and 656 utterances for the Innervillgraten dialect. As described by Toman et al. (2013b), this recording

[1] `http://sail.usc.edu/iemocap/`

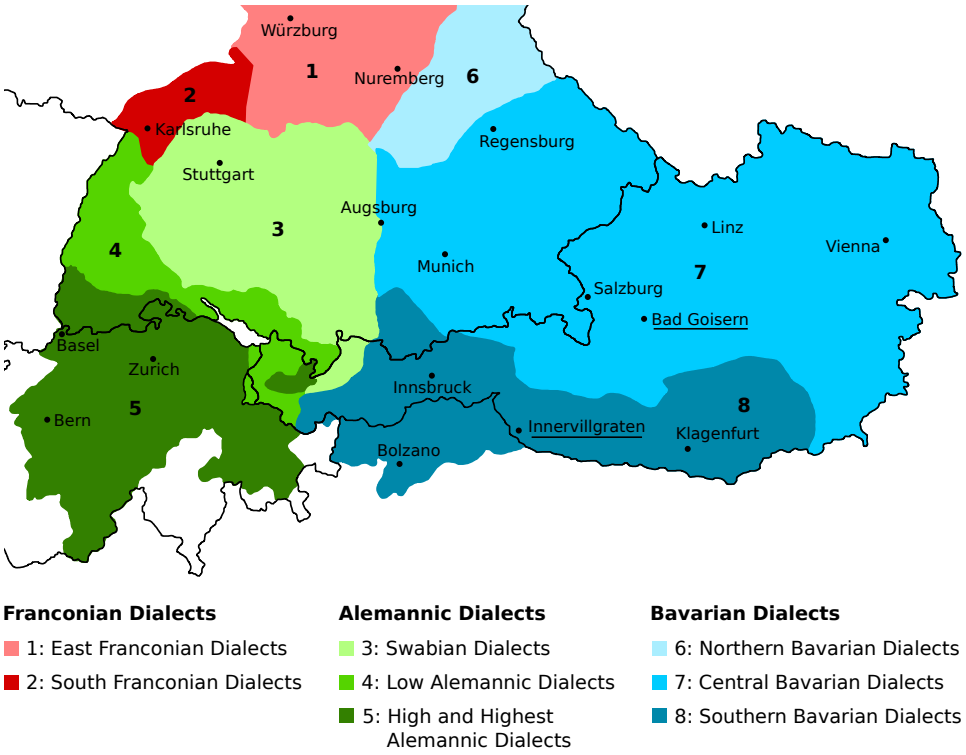



Figure 5.1: Geographic distribution of Upper German dialects, with the two towns of the recorded dialects highlighted. Image adapted from a public-domain image.[2]

script, as well as the corresponding phone set and transcriptions, were created in the following way: Per dialect, 18–20 hours of speech material of at least 10 speakers was collected by a phonetician in the respective town (no studio recording conditions). This material consisted of spontaneous speech elicited with given keywords, as well as translations from standard German given spontaneously by the dialect speakers. Next, this material was carefully analyzed phonetically and sentences were selected to be included in the recording script. All selected utterances were manually transcribed, thus providing a phone set and pronunciation dictionary, in addition to the transcriptions themselves.

Two female and two male speakers were selected for each dialect, based on fulfillment of the following linguistic criteria, which were assessed by a phonetician in the project team:

- Raised within the respective dialect (i.e., “native speaker”)
- Consistent application of characteristic phonological processes (e.g.,

[2] http://commons.wikimedia.org/wiki/File:Oberdeutsche_Dialekte.png

assimilations, deletions)

- Lexical knowledge and morpho-syntactic competence

These speakers were then recorded under the same conditions as described above for the project team members (motion capturing and audio under studio conditions). Because there are no defined orthographies for German dialects, producing dialect for read text is problematic, as the speakers would need to carry out a kind of translation from the written text to the spoken dialectal utterance. Therefore, a dialectal recording of each utterance was presented to the speakers as audio, while the words were displayed on a computer screen in standard German orthography.

Additionally, the same 223 utterances corpus of standard Austrian German as in the FMSC corpus were also recorded from the dialect speakers. The GIDS corpus of the eight dialect speakers will be released to the research community at a later point in time.

## 5.3 Multi-Modal Annotated Synchronous Corpus of Speech (MMASCS)

A detailed description of the third corpus, which includes also tongue motion recordings, was published by Schabus et al. (2014a), on which this section is largely based. The data of the MMASCS corpus is available under `http://schabus.xyz/phd/mmascs`.

This new corpus differs from existing ones in several aspects. Most importantly, it combines facial motion capture data with intra-oral EMA data. In comparison to optical motion capturing only, this has the obvious advantage of also providing tongue motion data, which is impossible to capture optically. In comparison to EMA data only, it has the advantage of providing a larger number of tracked points on the lips, eyelids, eyebrows and other areas of the face. While it is in principle possible to use EMA coils also on the face surface, the inexpensive and easy-to-attach optical markers are much less intrusive for the speaker than the EMA coils with their cable connection (one cable per coil) to the articulograph. Another difference is that our data is for Austrian German speech. One can imagine that it might be interesting to investigate inter-lingual differences in speech motion, once a larger number of corpora (of EMA and/or facial motion data) in various languages is available (of course speaker-specific effects would need to be accounted for). Finally, our data is different in that it comprises data of speech at three different speaking rates (normal, fast and slow).

In addition to general analytic usages, this corpus can be useful for other

fields of research. For example, in the context of 3D facial speech motion synthesis based on facial motion capturing data, where the additional tongue data can be used to train an additional synthesizer for tongue motion, similar to Beskow (2003). Cross-modality control models for speech synthesis, which have been investigated using EMA data and speech (Ling et al., 2008; Ling et al., 2009) and using facial motion data and speech (Hollenstein et al., 2013) can benefit from the usage of all three modalities in combination. Finally, we have used speech data at normal and fast speaking rates before to create ultra-fast synthetic speech via interpolation (Pucher et al., 2010a; Valentini-Botinhao et al., 2014). Incorporating additionally face and tongue motion data into such a system for ultra-fast speech might improve modeling and hence synthesis results.

### 5.3.1 Recordings

We have recorded a 30-year old male native speaker of Austrian German (the author of this dissertation) reading 320 phonetically diverse sentences off a computer screen. The recordings took place at the premises of Ludwig-Maximilians-Universität in Munich, inside a Studio Box Premium recording booth.[3] 223 sentences of the recording script are the same as for the FMSC corpus and the Standard Austrian German part of the GIDS corpus. The remaining 97 sentences were selected automatically from a large newspaper text collection based on the improvement caused when added to the selection with respect to the representation of the di-phone occurrence distribution in the entire large corpus.

As with the other corpora, facial movement was recorded using the OptiTrack system, as described in Chapter 4. Articulatory movement was recorded with a Carstens Medizinelektronik Articulograph AG501 EMA system.[4] In contrast to its predecessor AG500, the AG501 does not feature an acrylic glass cube around the speaker's head, which rendered simultaneous optical marker recording impossible. The AG501 produces alternating magnetic fields, thus inducing currents in the sensor coils attached to the speaker's tongue and mouth. The currents are transmitted via a cable from each sensor to the measurement unit, where they are measured and recorded. From these measurements, the system's software computes the 3D position of each sensor coil at 250 Hz. Articulatory sensors were placed on the back, middle and tip of the tongue, on the gums above the incisors and on the nasal bridge (all five on the mid-sagittal plane). Two more sensors were placed behind the ears, and finally an eighth sensor was placed on the lower lip, between the central lower lip and right lower lip markers of the OptiTrack

[3] http://www.acousticbooth-studiobox.com
[4] http://www.articulograph.de

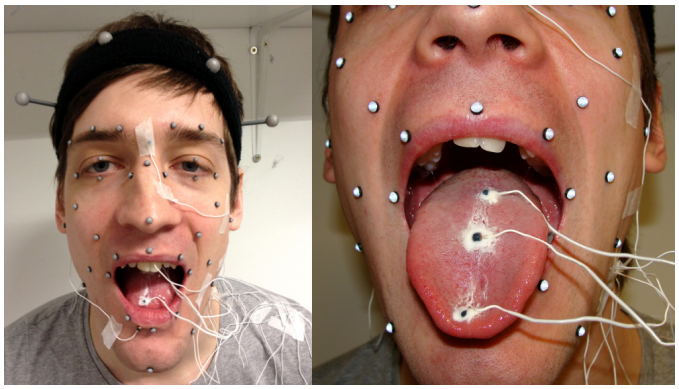


Figure 5.2: Placement of facial markers and EMA coils.

(a) (b)

Figure 5.3: Example frames from the gray scale video from the OptiTrack system (a) and the color video from the camcorder (b).

system. Figure 5.2 shows the position of most EMA sensors, as well as the facial markers for the optical system. Using the sensors on the nasal bridge, above the incisors and behind the ears, rigid head motion can be removed from the data. The EMA data was filtered using a finite impulse response low pass filter (Kaiser window) with cutoff frequencies of 40 Hz (tongue tip), 20 Hz (tongue middle, tongue back, lower lip), and 5 Hz (behind ears, upper incisors, nasal bridge).

Audio was recorded with a Sennheiser ME66 supercardioid microphone, with a John Hardy M1 pre-amplifier. The microphone signal as well as the synchronization signals from the EMA and OptiTrack systems were captured with a National Instruments Compact DAQ system at 25 600 Hz. Audio is encoded as 32-bit floating point PCM.

Additional video footage was recorded with a Sony DSR-PD100AP digital camcorder at 25 frames per second (50 fields interlaced) and from an al-

most frontal view. Figure 5.3 shows example frames from the two kinds of videos.

All 320 sentences were first recorded at a normal speaking rate, then again at a fast speaking rate and then again at a slow speaking rate, in direct succession with short breaks. Unfortunately, one of the tongue coils disengaged during the slow part, and the recordings had to be aborted after 130 slow sentences.

### 5.3.2 Release and Playback Software

For the release, the data has been synchronized and cut into separate files per utterance, in all modalities (audio, video, EMA data, facial movement data). Phone borders were determined by a flat-start forced alignment procedure using HTK (Young et al., 2006) and the resulting quin-phone full-context HTK label files and mono-phone label files are part of the release. Tracking errors, which are common in optical motion capturing (like marker swaps, trajectory gaps, etc.) have been manually corrected to a large extent. EMA data and facial marker data have been aligned in coordinate space based on the position of the markers on the nasal bridge of the two systems, after rigid head motion has been removed from both 3D data streams.

The facial motion and EMA data are provided in the form of text files containing matrices that represent spatial coordinates of markers/coils per row, with one column per time frame. Audio data is provided in the form of RIFF wave audio files, mono channel, 25 600 Hz, 16-bit signed integer PCM encoding. Video data is provided in the form of H264/AAC MPEG-4 video files.

The release also contains a playback software implemented in Python using OpenGL[5], which visualizes the 3D data (facial markers and tongue coils) and also simultaneously plays back the corresponding audio. Figure 5.4 shows a screen shot of this software.

### 5.3.3 Data Analysis

To get a better understanding of the data we have recorded, we performed a statistical analysis of the corpus. As already mentioned, we have 320 sentences for both normal and fast speaking rate, and 130 sentences for slow speaking rate. For symmetry, all analytics in this section are based on the 130 sentences which we have available in all three speaking rates.

[5] http://www.opengl.org

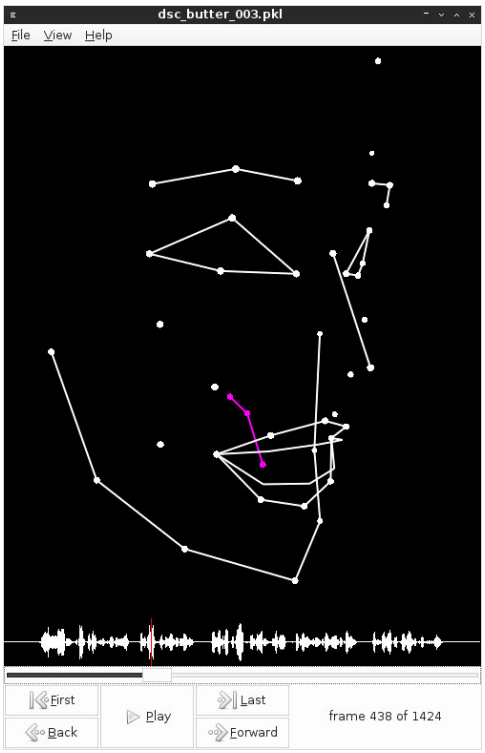



Figure 5.4: Screen shot of 3D data visualization software included in the corpus release.

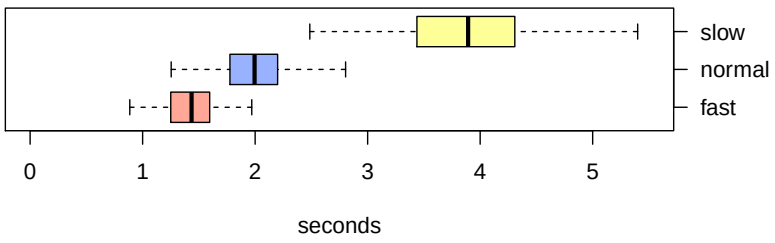



Figure 5.5: Boxplots of utterance durations for the three speaking rates (outliers not shown).

Figure 5.5 shows the distributions of the utterance durations for the three speaking rates as boxplots, disregarding initial and terminal silences and intra-utterance pauses. As the same 130 sentences were used, the figure shows that there is a significant difference in duration between the three speaking rates.

To quantify the different speaking rates in more depth, we have looked at the phone durations as determined by the flat-start forced alignment procedure. Figure 5.6 shows boxplots of the phone durations, excluding all silences and

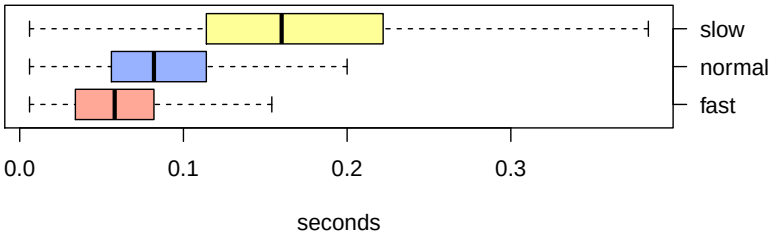



Figure 5.6: Boxplots of phone durations for the three speaking rates (outliers not shown).

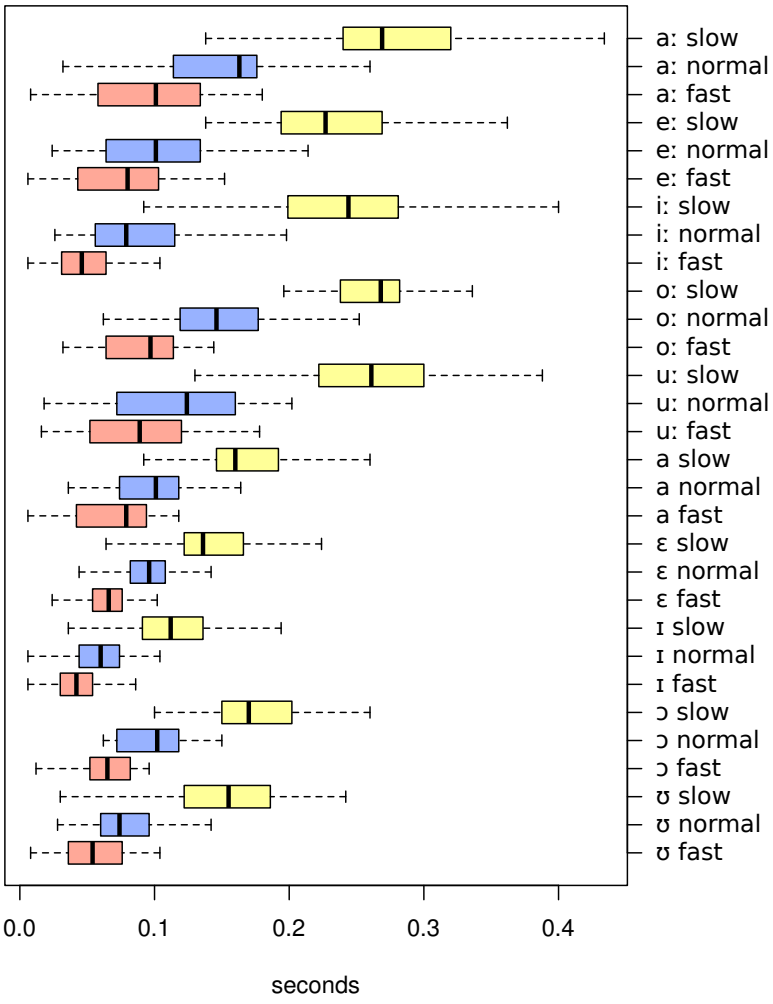



Figure 5.7: Boxplots of phone durations of some common short and long vowels, for the three speaking rates (outliers not shown).

pauses. The median phone durations for the slow, normal and fast speaking rate data are 160 ms, 82 ms and 58 ms, respectively, which are equivalent to 375, 732 and 1034 phones per minute, respectively. In addition to the occurrence of longer phones, the data also show a larger variability in phone duration with decreasing speaking rate. Furthermore, when we partition this data by phone, we can observe that the change of duration between speaking rates is larger for long vowels and diphthongs than for short vowels and stops. This is illustrated in Figure 5.7, which shows the duration distributions for some common short and long vowels.

To achieve a faster speaking rate, i.e., to articulate the same sequence of phones in a shorter time, three factors can be modified: 1) the velocity of the articulator movements can be increased, 2) the distance between the target articulator positions can be reduced, and/or 3) the duration of phases with stable articulator position can be shortened. Given the data of our corpus, the first two are straightforward to assess, and shall be investigated in the following.

Regarding the first factor, we have computed the movement velocities for the three tongue sensors based on the distance traveled between every two consecutive frames of the EMA trajectories. Figure 5.8 shows the distributions of peak velocities (greatest velocity within a phone) for the three speaking rates. Although this data may contain some noise, the increase in tongue motion velocity from slow to normal and from normal to fast speaking rate is clearly visible. The same data, but partitioned by phone, is shown in Figure 5.9. Again, this data is not completely reliable due to possible problems in the automatic alignment and possible tracking errors, and due to the fact that some phones do not occur very often in the corpus. Nevertheless, it is interesting to see that the order of phones is quite similar across the three speaking rates when sorted by median (as in Figure 5.9). In particular, phones near the close/front corner of the IPA vowel chart ([i], [iː], [y], [yː], [ɪ], [ʏ], [eː], [øː]) and certain fricatives ([s], [ʃ], [ç]) exhibit low peak velocities (and thus appear close to the bottom of Figure 5.9), whereas vowels far from the close/front corner ([uː], [ʊ], [oː], [ɔ], [a]) and diphthongs exhibit high peak velocities (and thus appear close to the top of Figure 5.9).

Regarding the second factor, i.e., the influence of speaking rate on tongue target positions, we have gathered for each of the three tongue sensors (back, middle and tip of the tongue) the deviation from its average position, as shown in the boxplots of Figure 5.10. The figure shows each of the $x$, $y$ and $z$ coordinates separately, which correspond to the left/right, up/down and front/back directions from the speaker's point of view. A slight decrease in positional variability can be seen for increased speaking rate, suggesting that tongue movement needs to be reduced for faster speech.

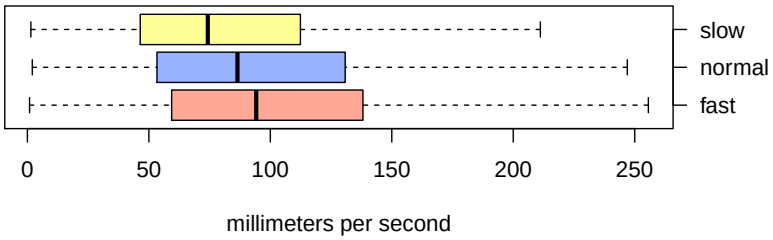
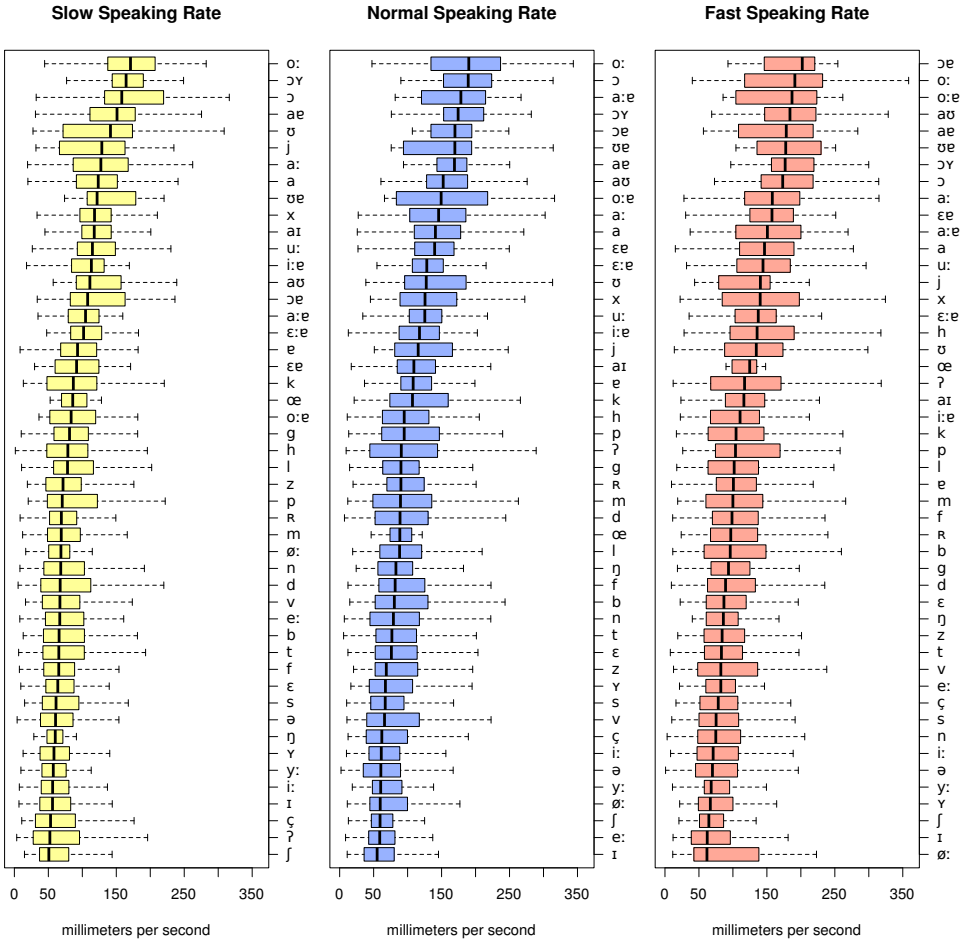



Figure 5.8: Boxplots of peak movement velocities of the three tongue sensors (outliers not shown).




(a) (b) (c)

Figure 5.9: Boxplots of peak movement velocities of the three tongue sensors per phone (outliers not shown), for (a) slow, (b) normal, and (c) fast speaking rates.

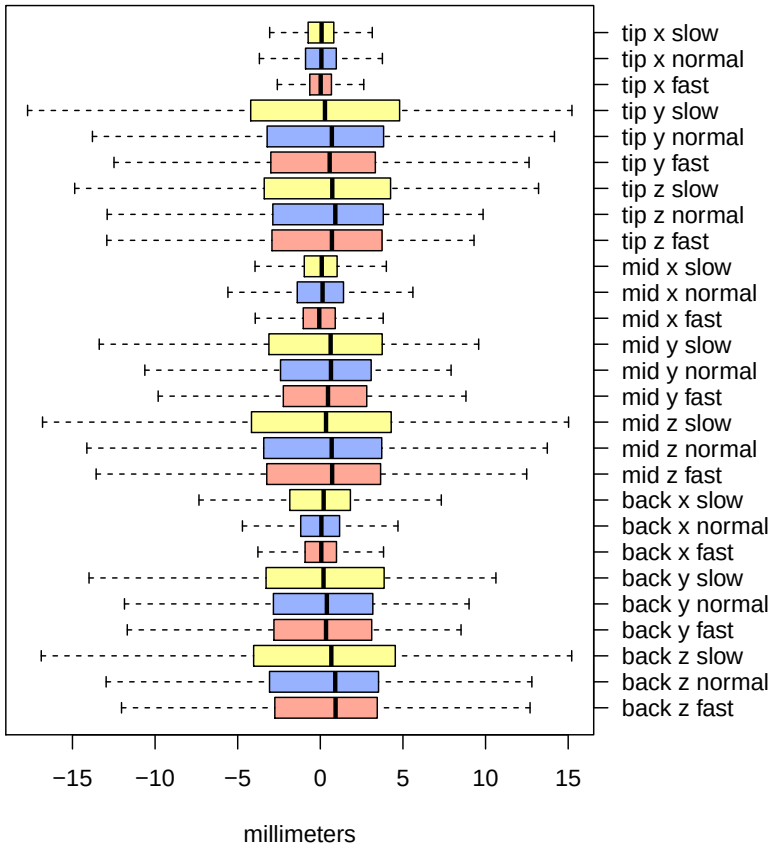



Figure 5.10: Boxplots of mean-normalized spatial coordinates of the three tongue sensors (outliers not shown).

These findings are in line with, e.g., Flege (1988), where increased speaking rate is reported to result from a combination of both increased movement velocity and decreased divergence of the tongue from a "centroid" or "rest" position.

Similar to Yehia et al. (1998); Jiang et al. (2000); Beskow (2003), we have also looked at how well the tongue motion data can be predicted from the facial motion data and vice versa. In a 10-fold cross validation setup, we have computed a linear regression to predict one tenth of the tongue (face) data from the corresponding face (tongue) data, where the other nine tenths of the data are used to estimate the predictor. Then Pearson's correlation coefficients are computed between the predicted and the originally recorded tongue (face) data. Note that we excluded the face markers on the eyebrows and eyelids for this analysis step because their movement can be expected to be unrelated to phone articulation. The average correlation coefficients resulting from this procedure are shown in Table 5.1. The results are com-

Table 5.1: Average Pearson's correlation coefficients between measured and predicted marker coordinates.

| Speaking Rate | Face from Tongue | Tongue from Face |
|---|---|---|
| Slow | 0.234 | 0.445 |
| Normal | 0.226 | 0.558 |
| Fast | 0.279 | 0.523 |

Table 5.2: Number of recorded utterances per corpus, speaker and variety: standard Austrian German (SAG), Bad Goisern dialect (GOI), Innvervillgraten dialect (IVG).

| Corpus Name | Speaker | Gender | #SAG | #GOI | #IVG | speaking rate |
|---|---|---|---|---|---|---|
| FMSC | SF1 | female | 223 | | | normal |
| FMSC | SM1 | male | 223 | | | normal |
| FMSC | SM2 | male | 223 | | | normal |
| GIDS | GF1 | female | 223 | 665 | | normal |
| GIDS | GF2 | female | 223 | 665 | | normal |
| GIDS | GM1 | male | 223 | 665 | | normal |
| GIDS | GM2 | male | 223 | 665 | | normal |
| GIDS | IF1 | female | 223 | | 656 | normal |
| GIDS | IF2 | female | 223 | | 656 | normal |
| GIDS | IM1 | male | 223 | | 656 | normal |
| GIDS | IM2 | male | 223 | | 656 | normal |
| MMASCS | SM2 | male | 223 | | | normal |
| MMASCS | SM2 | male | 223 | | | fast |
| MMASCS | SM2 | male | 130 | | | slow |

parable to the ones of the "Sentences, 3 coils" condition[6] of Beskow (2003) (tongue from face: 0.525, face from tongue: 0.357), which is the condition most similar to our setup. It can be seen that prediction of tongue motion from face motion is more successful than prediction in the opposite direction. There does not seem to be a clear influence of speaking rate on the predictability of tongue motion from face motion and vice versa.

To conclude this chapter on the data collections produced for this dissertation, Table 5.2 gives an overview of all recorded data of the three corpora.

[6] Using sentence recordings rather than nonsense vowel-consonant-vowel and consonant-vowel-consonant utterances; using 3 tongue coils only rather than including also jaw and lip coils.

# Chapter 6

# Synchronization of Speech and Motion

This chapter investigates *joint audiovisual* HSMMs as a way of addressing the problem of synchronization between the generated acoustic and visual speech. This chapter is closely related to an earlier publication (Schabus et al., 2014a). Different acoustic, visual, and joint audiovisual models for four different Austrian German speakers were trained and we show that the joint models perform better compared to other approaches in terms of synchronization quality of the synthesized visual speech. In addition, a detailed analysis of the acoustic and visual alignment is provided for the different models. Importantly, the joint audiovisual modeling does not decrease the acoustic synthetic speech quality compared to acoustic-only modeling so that there is a clear advantage in the common duration model of the joint audiovisual modeling approach that is used for synchronizing acoustic and visual parameter sequences.

## 6.1 Introduction

Chapter 2 presented existing approaches/systems for audiovisual speech synthesis, making a distinction between image-based methods and methods using 3D head models. The work presented in this dissertation belongs to the latter category, as described in Chapter 4. Regardless of the way the visual speech signal is captured and represented, the two signals generated by an audiovisual speech synthesis system need to be correctly synchronized in order to deliver a believable experience of bi-modal speech to the user, as already briefly discussed at the end of Chapter 2. To this end, this chapter proposes a joint audiovisual HSMM-based approach, where audible speech

and facial motion are combined into a single bi-modal model.

In statistical data-driven audiovisual synthesis, commonly separate acoustic and visual models are trained (L. Wang et al., 2011b; Sako et al., 2000; Masuko et al., 1998; Tamura et al., 1998a; Hofer and Richmond, 2010; Hofer et al., 2008), sometimes together with an additional explicit time difference model to correctly synchronize the two modalities (Govokhina et al., 2007; Bailly et al., 2009). In contrast, we propose to train one joint audiovisual model (with acoustic and visual streams), such that the likelihood of the model generating the training data is maximized globally, across the two modalities, during model parameter estimation. This results in a single duration model used for both modalities, thus eliminating the need for additional synchronization measures. In this way, we intend to create simple and direct models for audiovisual speech synthesis, which can cope with most effects of co-articulation and inter-modal asynchrony naturally through five-state quin-phone full-context modeling. Bailly et al., 2009 also argue that states can capture some inter-modal asynchrony since transient and stable parts of the trajectories of different modalities need not necessarily be modeled by the same state, and that multi-phone context models can capture co-articulation effects. Notably, an early work on audiovisual HMM-synthesis (Tamura et al., 1999) also applied joint modeling in our sense, however without investigating its benefits in detail. Also, the current HMM-modeling techniques and high-fidelity visual parameter acquisition we use distinguish our work from Tamura et al., 1999.

Therefore the main purpose of this chapter is to investigate whether the proposed joint audiovisual modeling approach provides clear improvements over separate audio and visual modeling. We argue that the main weakness of separate modeling stems from the difficulty to capture (and even define) clear temporal unit borders for the visual modality. Our analysis shows that visual-only training yields models which fail to find suitable borders for some phones when we carry out forced alignment on our training data. An explicit audio/video lag model used for modality synchronization, which is trained on such borders (as by Govokhina et al., 2007 and Bailly et al., 2009) might still suffer from these problems, even if the borders in the training data are hand-labeled (as done by Terry, 2011). Furthermore, the quality of the synthesized trajectories themselves can be expected to degrade if observation assignment to units is unclear during training.

On the other hand, there are situations where the targets to which speech needs to be synchronized are much clearer, like singing synthesis (Saino et al., 2006), where explicit lag models have been used successfully for synchronizing speech to sheet music (in that case, the sheet music defines fixed and exact synchronization target points in time).

By using the system pipeline described in Chapter 4, we employ a state-of-

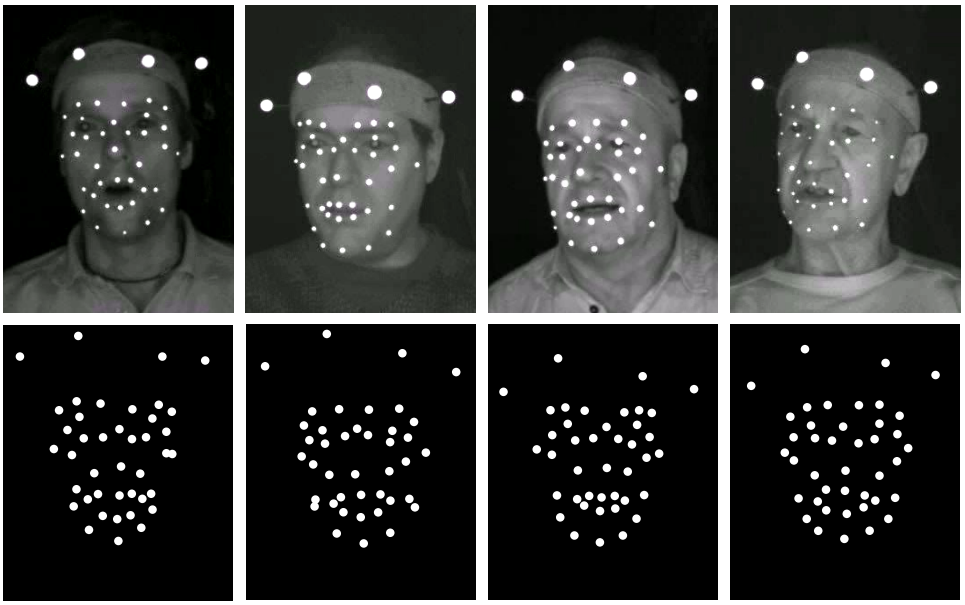


Figure 6.1: Frames from grayscale videos showing facial marker layout (top) for four different speakers and corresponding renderings of 3D marker data (bottom).

the-art HSMM modeling framework and we use current animation-industry-standard motion tracking and character animation technology for the visual modality. In this regard our work differs strongly from conceptually related previous work (Sako et al., 2000; Masuko et al., 1998; Tamura et al., 1998a).

## 6.2 System Description

For the investigations in this chapter, the recordings in Standard Austrian German from the four male dialect speakers of the GIDS corpus (cf. Chapter 5) were used as training data, which amounts to 223 utterances and roughly 11 minutes total per speaker. Global head motion removal, positional normalization and PCA-based visual feature extraction were carried out as discussed in Chapter 4, using $k = 30$ dimensions for the visual signal. Figure 6.1 shows example frames for the four speakers.

For training regular audio speech models, we used the CSTR/EMIME TTS system training scripts (Yamagishi and Watts, 2010) and HTS version 2.1 to train context-dependent, five-state, left-to-right, MSD-HSMMs (Zen et al., 2007d). As audio features we used 39+1 mel-cepstral features, log F0 and 25 band-limited aperiodicity measures, extracted from 44.1 kHz speech, as

it is done in the CSTR/EMIME system. Speech signals are re-synthesized from these features using the STRAIGHT vocoder (Kawahara et al., 1999). All features are augmented by their dynamic features ($\Delta$ and $\Delta^2$) (Tokuda et al., 1995). For each of the three audio features, the models are clustered separately state-wise by means of decision-tree based context clustering using linguistically motivated questions on the phonetic, segmental, syllable, word and utterance levels. State durations are modeled explicitly rather than via state transition probabilities (HSMMs rather than HMMs, Zen et al., 2004), and duration models are also clustered using a single decision-tree across all five states. The feature questions used for the clustering are based on the English question set in the EMIME system (Yamagishi and Watts, 2010) with adaptations towards our German phone set. They are listed by Pucher et al. (2010b), except that we do not use multiple dialects here and that we also included the Phoneme Equivalence Class (PEC)/viseme classes of preceding, current, and succeeding phones (as described below).

In short, for *audio-only* modeling, we apply the state-of-the-art CSTR/EMIME HTS system without modifications. For *visual-only* modeling, we use the same system but with only one feature stream for the visual PCA-space features. In order to obtain the same frame rate as the audio features (5ms frame shift, i.e., 200 frames per second), we have up-sampled (interpolated) the visual features from their native 100 frames to 200 frames per second. Similar to the mel-cepstral features, they are also augmented by their dynamic features and the models are clustered using the same set of questions. This results in a speaker-dependent text-to-visual speech system. Furthermore, for *joint audiovisual* modeling, we merge the two into a system that trains models for the three audio features (mel-cepstral, F0, aperiodicity) and the visual features simultaneously. This is achieved by adding an additional stream to the audio-only system, with separate state-wise clustering. The structure of the audio, visual and audiovisual systems is shown in Figure 4.10 in Chapter 4, where the black and blue parts constitute the audio-only system, the black and red parts constitute the visual-only system and all parts constitute the joint audiovisual system.

As discussed in Chapter 4, the generated visual parameter sequences can be re-projected into the full-dimensional space, thus turning the features sequences into sequences of facial marker coordinates again. Using a retargeting function for a 3D head model, these marker coordinate sequences can be turned into facial animations.

As we have added an additional non-standard feature to the well-established HSMM training system, it is of interest to see how the new feature is handled by the system. One potentially informative parameter for this is the size of the clustering trees. Table 6.1 gives the number of leaf nodes, i.e., the number of distinct observation probability density functions, resulting from

Table 6.1: Average number (across four speakers) of leaf nodes in the clustering trees after training.

| | | State | | | | | |
|---|---|---|---|---|---|---|---|
| Training | Feature | 1 | 2 | 3 | 4 | 5 | Total |
| Audio | Mel-cepstral | 58 | 61 | 69 | 66 | 67 | 320 |
| | Log F0 | 146 | 219 | 241 | 149 | 100 | 856 |
| | Band-Ap | 27 | 34 | 36 | 30 | 25 | 152 |
| | Duration | | | | | | 163 |
| Audiovisual | Mel-cepstral | 57 | 63 | 67 | 58 | 61 | 306 |
| | Log F0 | 164 | 218 | 259 | 164 | 121 | 925 |
| | Band-Ap | 27 | 31 | 32 | 23 | 27 | 140 |
| | Visual | 258 | 526 | 551 | 417 | 291 | 2042 |
| | Duration | | | | | | 208 |
| Visual | Visual | 354 | 504 | 418 | 345 | 314 | 1934 |
| | Duration | | | | | | 312 |

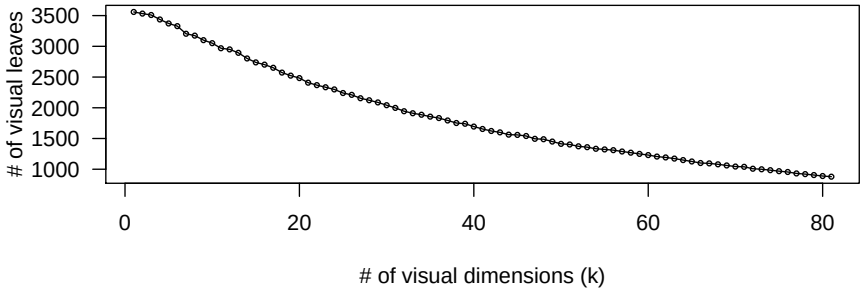



Figure 6.2: Average number (across four speakers) of total leaf nodes in the visual clustering trees as a function of visual PCA dimensions kept ($k$).

the audio, audiovisual and visual training procedures, averaged across the four speakers. The absolute numbers in such a table of course grow with the size of the training corpus, but we can observe that the trees for the visual features are substantially bigger than the ones for the other features, which is still true if we choose a different dimensionality $k$ to represent our visual data, as illustrated in Figure 6.2 where the number of visual leaf nodes is shown as a function of $k$ resulting from audiovisual training. This is somewhat surprising, given that the visual parameter trajectories appear to be quite smooth in general (see Figure 6.4 for an example). We interpret this as a strong dependency on context of our visual data.

We also find that the size of the duration tree of the visual-only voice model is roughly twice the size of the audio-only duration tree, and that in the combined audiovisual system we also see an (albeit smaller) increase in size of the duration tree. Duration and audiovisual synchronization will be dis-

cussed in more detail in Sections 6.3 and 6.4, but we can already see from these numbers that duration modeling for the visual features seems to work differently from the audio features.

In many approaches to (audio-)visual speech processing, the concept of *visemes* (Fisher, 1968; Chen, 2001; Massaro et al., 2012) or, more generally, PECs (Bernstein, 2012) is used. The idea is roughly that phone(me)s which have similar or even indistinguishable visual appearance (but which may still be very different in acoustic terms) are grouped together for visual modeling. It is easy to integrate this concept into the HSMM modeling framework, even with the flexibility to use the concept only partially: By "offering" to the model clustering algorithm additional questions that correspond to such groupings of phones according to their visual properties, the maximum description length criterion (cf. Section 3.3.4) will automatically make use of such PEC questions when and only when they are useful. To determine to what degree PECs are beneficial or even necessary for visual speech modeling in our setting, it is therefore sufficient to simply provide additional questions alongside the ones mentioned earlier (e.g., phones and phone groups based on acoustic criteria) and then to see whether these are used to cluster the data at hand.

Based on the "easy set" in by Bernstein (2012), with adaptations towards our phone set for German, we have added the following six PECs as possible clustering questions: {p, b, m}, {f, v}, {t, d, s, z}, {k, g, n, ŋ, l, h, j, ç, x}, {oː, uː, yː, øː}, {ɔ, ʊ, ʏ, œ}.

Assuming that such PECs are useful for modeling the visual features but not the acoustic ones, these questions should appear often in the clustering trees for the former and rarely (or not at all) for the latter, when they are "offered" at all clustering steps of all features. The percentages of decision tree leaves affected by PEC questions are given in Table 6.2 for the three training procedures and all features, averaged across four speakers. Here we consider a leaf "affected" if at least one PEC question was answered affirmatively on the path from the root to the leaf. In line with the expectations mentioned before, we see that PEC questions clearly play a more important role in clustering the models for the visual features than for the acoustic ones, although they are also used for the latter to some extent. PEC questions are especially relevant for the third (22.3%) and fourth (26.2%) states of the visual stream. Interestingly, the presence of the visual features also has an impact on the duration clustering in this respect (in addition to making the duration trees larger, as we have discussed earlier): The duration trees of the visual-only and the audiovisual models contain a higher percentage of PEC-affected leaves than the acoustic-only models.

We conclude from these findings that the addition of clustering questions specifically targeted towards visual features such as visemes or PECs can be

Table 6.2: Average percentage (across four speakers) of leaf nodes affected by PEC questions.

| | | State | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Feature | 1 | 2 | 3 | 4 | 5 | Overall |
| Audio | Mel-cepstral | 8.9 | 6.1 | 5.8 | 5.4 | 7.6 | 6.7 |
| | Log F0 | 7.9 | 5.9 | 4.9 | 3.3 | 4.9 | 5.4 |
| | Band-Ap | 1.0 | 4.2 | 3.9 | 2.6 | 0.0 | 2.5 |
| | Duration | | | | | | 5.1 |
| Audiovisual | Mel-cepstral | 9.3 | 4.8 | 4.6 | 1.7 | 5.4 | 5.1 |
| | Log F0 | 8.1 | 7.4 | 7.0 | 6.0 | 6.2 | 7.0 |
| | Band-Ap | 6.1 | 3.3 | 0.7 | 2.2 | 2.7 | 2.9 |
| | Visual | 13.1 | 10.2 | 22.3 | 26.2 | 13.5 | 17.3 |
| | Duration | | | | | | 12.7 |
| Visual | Visual | 13.9 | 26.4 | 22.9 | 26.7 | 17.5 | 21.9 |
| | Duration | | | | | | 13.1 |

helpful in modeling the visual modality in this framework.

## 6.3 Audiovisual Synchronization Strategies

To achieve the goal of text-to-audiovisual-speech synthesis, both an acoustic speech signal and a visual speech signal (animation) need to be created given some input text, and in addition to being natural or believable individually, the two generated sequences need to *match temporally*. With the three trained models described in the previous section available (audio-only, visual-only and joint-audiovisual, each with its own duration model), there are several possible strategies that lead to a combined audiovisual sequence generated for some new input text. The following six subsections describe five different ways of combining the two uni-modal models, as well as the sixth "synchronization strategy" emerging directly from using a bi-modal audiovisual model.

### 6.3.1 Unsynchronized (*unsync*)

The simplest strategy using the separately trained models is to synthesize from each model independently and then just add the two generated sequences together. This has the advantage that each model will generate its sequence "naturally", i.e., the way that directly emerges from the training process of the respective model. An important disadvantage is that there are no synchronization constraints whatsoever, and the total length of the

generated audio and visual sequences may even differ. We will refer to this method, which uses two duration models, as *unsync* for short.

### 6.3.2 Audio Utterance Length (*uttlen-audio*)

While still using both duration models, we can ensure equal sequence length by adjusting the speaking rate parameter $\rho$ in the synthesis step (Yoshimura et al., 1998). The state durations of an utterance consisting of $K$ states (i.e., $K/5$ phones) are given by

$$d_A(k) = \mu_A(k) + \rho \cdot \sigma_A^2(k) \quad \text{for } 1 \leq k \leq K, \tag{6.1}$$

where $\mu_A(k)$ and $\sigma_A^2(k)$ denote the mean and variance of the audio duration model for state $k$, respectively. When $\rho$ is set to 0 for synthesis, we obtain speech in average speaking rate, with $\rho < 0$ we obtain faster and with $\rho > 0$ slower speech. We can synthesize acoustically without constraints ($\rho_A = 0$), and then determine the $\rho_V$ required for visual synthesis that will yield the same utterance length:

$$D_A = \sum_{k=1}^{K} d_A(k) = \sum_{k=1}^{K} \mu_A(k), \tag{6.2}$$

$$\rho_V = \frac{D_A - \sum_{k=1}^{K} \mu_V(k)}{\sum_{k=1}^{K} \sigma_V^2(k)}, \tag{6.3}$$

where $\mu_V(k)$ and $\sigma_V^2(k)$ denote the mean and variance of the visual duration model for state $K$.

This will produce an audio and visual parameter sequence for the utterance which are exactly of the same length, but still each use their respective duration model. We will refer to this strategy, which exhibits the "natural" audio duration, as *uttlen-audio* for short.

### 6.3.3 Visual Utterance Length (*uttlen-visual*)

Symmetrically, by flipping the roles of audio and visual models, we obtain another strategy that exhibits the "natural" visual duration, referred to as *uttlen-visual*.

### 6.3.4 Copy Audio Duration (*durcopy-audio*)

In order to achieve tighter synchronization on the phone level, we can decide to use only one of the two duration models, e.g., the audio duration model

Table 6.3: Synchronization strategies for audiovisual synthesis.

| Name | Description |
|---|---|
| unsync | unsynchronized separate duration models |
| uttlen-audio | utterance length determined by audio duration model |
| uttlen-visual | utterance length determined by visual duration model |
| durcopy-audio | audio duration model used for both modalities |
| durcopy-visual | visual duration model used for both modalities |
| audiovisual | features trained jointly, audiovisual duration model |

for both audio and visual synthesis. This is equivalent to replacing the visual duration models and trees with the ones obtained from audio training. The advantage here is the tighter synchronization, a possible disadvantage is that a new duration model is forced upon the visual system which might not match the visual feature models. We will refer to this strategy as *durcopy-audio*.

### 6.3.5 Copy Visual Duration (*durcopy-visual*)

Likewise, we can replace the audio duration model with the visual one, which we will call *durcopy-visual*.

### 6.3.6 Joint Audiovisual (*audiovisual*)

Finally, the audiovisual voice model with jointly trained features and with a single audiovisual duration model generates synchronized parameter trajectories implicitly. A priori it is not clear what kind of effect the additional visual stream will have on the quality of the generated audio samples. One can imagine that the additional information will lead to more robust parameter estimation and thus to an improvement of audio quality. On the other hand, if the two signals reveal themselves to be rather inconsistent, a negative effect on audio quality could arise. We will refer to this strategy as *audiovisual*.

The six synchronization strategies are summarized in Table 6.3. Note that the first three (*unsync*, *uttlen-audio*, *uttlen-visual*) use two duration models whereas the last three (*durcopy-audio*, *durcopy-visual*, *audiovisual*) each use a different single duration model. Furthermore note that *unsync*, *uttlen-audio* and *durcopy-audio* produce synthetic speech identical to what the regular audio-only system would produce.

The *unsync* method does not guarantee that audio and visual sequences have the same length, but since both models are trained on the same syn-

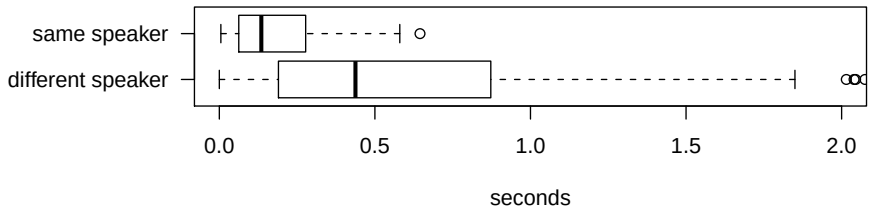



Figure 6.3: Boxplots for the differences in utterance length between audio-only and visual-only synthesized utterances. For 23 test utterances and 4 speakers, the top boxplot contains all 92 combinations where audio and visual models were from the same speaker, and the bottom boxplot contains all 276 combinations where the sequences were synthesized using two different speakers' models.

chronous corpus, the deviation can be expected to be small, as illustrated in Figure 6.3, which shows boxplots of the difference in length when the same utterance is synthesized from an audio-only and from a visual-only model separately. The figure also shows clearly that this difference is significantly smaller when the two models are from the same speaker (and thus trained on a synchronous corpus), suggesting that this synchronization strategy can not work for mixed-speaker setups, if at all.

The *durcopy-audio* method is a straightforward choice to align the borders of both sequences by simply using the borders predicted by the audio model also for the visual model, applied for example by Sako et al. (2000) and Schabus et al. (2011).

The *uttlen-audio* method is interestingly similar to the explicit lag models of Govokhina et al. (2007) and Bailly et al. (2009): with *uttlen-audio*, audio is synthesized independently of the visual features, and a separate visual duration model predicts the visual phone borders, while the length constraint ensures equal total length of the two sequences. The separate visual model results from several iterations of embedded training on visual-only data. The main difference is that Govokhina et al. (2007) and Bailly et al. (2009) predict the visual phone borders as a relative offset to the audio borders, where the offsets are iteratively re-estimated based on visual forced alignment.

## 6.4 Alignment Analysis

This section analyzes the temporal alignment behavior of the different models described in the previous section. Although speech movements and the resulting sounds are synchronous in general, it is not clear a priori whether

the borders between phones in the visual speech signal should be the same as in the audio speech signal. For example, at the beginning of an utterance, anticipatory gestures can begin in the speech movement signal well before any audible sound is produced. Although somewhat unnatural, it is commonplace in audio speech synthesis (as well as speech recognition) to define sharp borders between the phones of an utterance and to compensate for co-articulation effects by employing context-dependent modeling strategies (as it is also done in the HTS system we use). Given an acoustic model, such phone borders can be found automatically by forced alignment of the known phone sequence to some speech data.

We have applied HSMM-based forced alignment via the *HSMMAlign* tool from HTS version 2.2 to our training data using the different models we have trained, in order to understand the temporal differences between auditory, visual and joint audiovisual modeling. Given the auditory model and the auditory data, this produces for each of the 200 utterances in the training corpus the most likely phone borders that would make the auditory model generate the speech parameters of this utterance. Likewise for the visual model and data, as well as the audiovisual model and data.

Figure 6.4 shows an example sentence with the corresponding forced alignment results. In the first row, the visual-only model was used to align the visual data, the resulting phone borders are designated by black vertical lines. For easier interpretation, the plot shows the Euclidean distances between the central upper lip and central lower lip markers as well as between the left and right mouth corner makers, instead of PCA components. In the third row, the auditory-only model was used to align the auditory data. Here, the first three mel-cepstral features are drawn in red in decreasing thickness and F0 is drawn in green. The low flat portions of the F0 signal represent unvoiced parts (undefined F0). All features have been re-scaled to fit into the same vertical range. The second row combines all features, and the alignment was determined using the joint audiovisual model. The bottom row shows the spectrogram of the utterance. It is apparent that there is a difference between the three resulting alignments.

In order to quantify this temporal alignment difference between the three models, we have computed the alignments for all 200 utterances for all four speakers. Then, to assess the degree of agreement between any two models, we have computed the time percentage of each utterance where the two alignments agree. For an utterance $u$ consisting of the phone sequence $(p_1, p_2, ..., p_n)$, we compute the agreement percentage $M(u, A, B)$ between

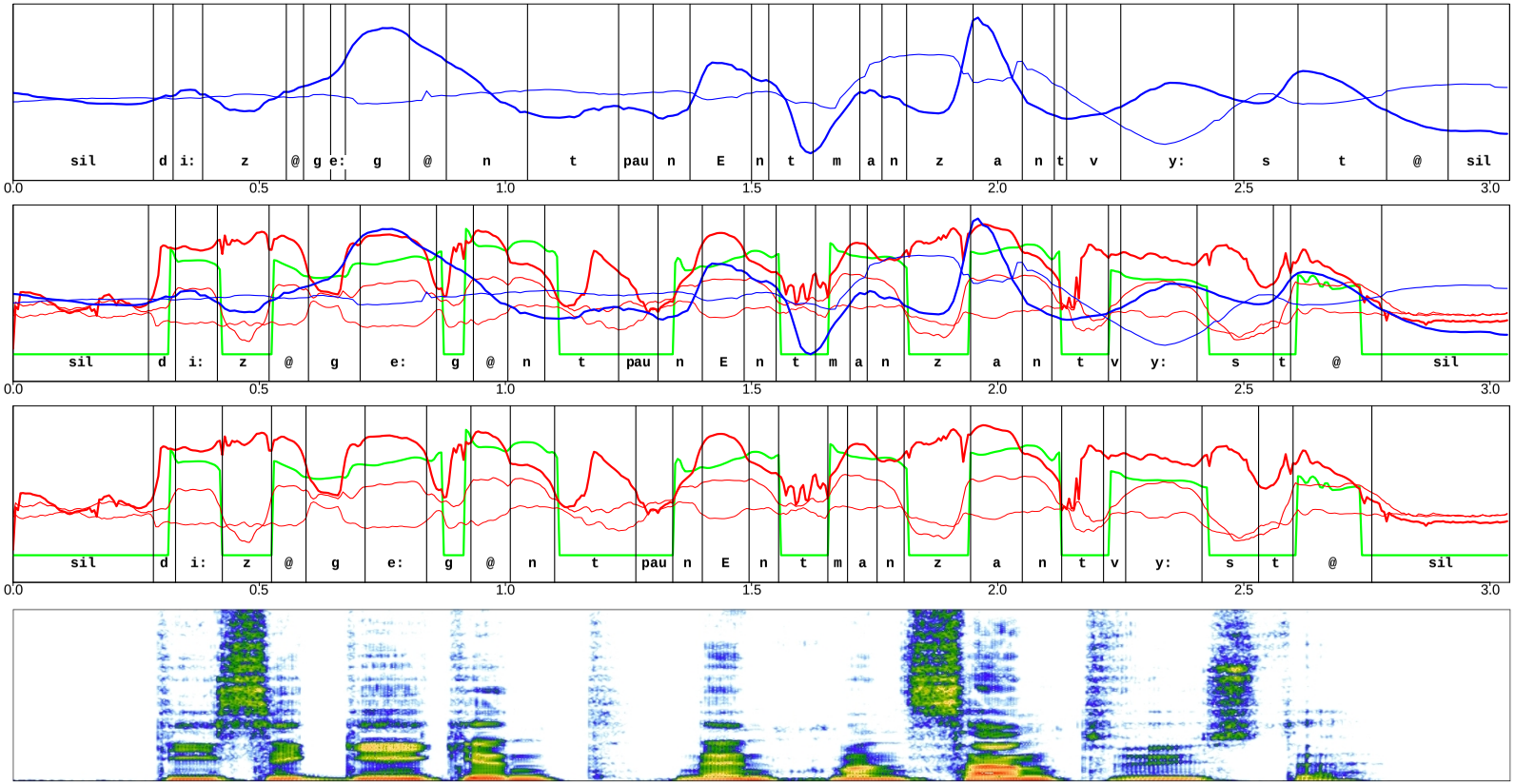



Figure 6.4: Result of forced alignment using visual (first row), audiovisual (second row) and audio (third row) models and data. The plots show the Euclidean distance between the central upper lip and central lower lip markers (thick blue line), the Euclidean distance between the left and right mouth corner markers (thin blue line), the first three MFCCs (red, with decreasing thickness) and F0 (green). The different features have been re-scaled to fit into the same vertical range. The bottom row shows the spectrogram. The sentence is "Diese Gegend nennt man Sandwüste" (this area is called a sand desert).

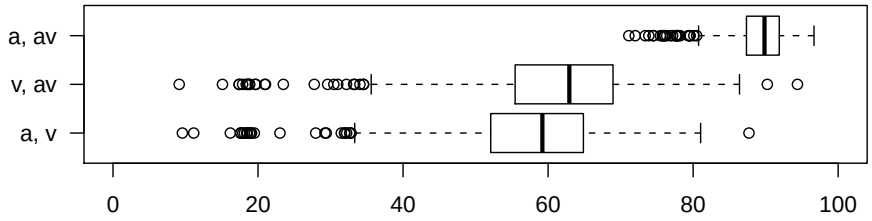



Figure 6.5: Boxplots for matching percentage per utterance ($M_{utt}$) for audio and audiovisual models (a, av), visual and audiovisual models (v, av) and audio and visual models (a, v).

two models $A, B \in \{audio, visual, audiovisual\}$ for that utterance as

$$M_{utt}(u, A, B) = \frac{100}{e_{p_n,A}} \cdot \sum_{i=1}^{n} \max(0, \min(e_{p_i,A}, e_{p_i,B}) - \max(b_{p_i,A}, b_{p_i,B})), \tag{6.4}$$

where $b_{p_i,X}$ and $e_{p_i,X}$ denote the beginning and the end of phone $p_i$ as determined by *HSMMAlign* using model $X$. Note that $e_{p_n,A} = e_{p_n,B}$ is simply the total length of the utterance.

The resulting matching percentages of all 800 utterances are shown as boxplots in Figure 6.5. The degree of agreement between the auditory and the audiovisual models is much higher (median 89.84%) than between the visual and audiovisual models (median 62.93%) and between the auditory and visual models (median 59.21%). The utterance in Figure 6.4 is a typical example in this regard with (a, av)-match 89.31%, (v, av)-match 62.66%, and (a, v)-match 58.88%.

We have also computed the matching percentages for any two methods for each individual phone. The percentage is calculated as the amount of time that both alignments consider as being part of the phone divided by the average of the two phone lengths, formally

$$M_{phone}(p_i, A, B) = \frac{100 \cdot \max(0, \min(e_{p_i,A}, e_{p_i,B}) - \max(b_{p_i,A}, b_{p_i,B}))}{\frac{1}{2}((e_{p_i,A} - b_{p_i,A}) + (e_{p_i,B} - b_{p_i,B}))}. \tag{6.5}$$

Figure 6.6 shows the results grouped by phones (i.e., central phones of the respective quin-phone full-contexts). Apart from the overall better match between auditory and audiovisual (Figure 6.6a) compared to the two other pairs (Figure 6.6b and 6.6c), which is also shown by Figure 6.5, it can be seen in these plots that the bottom 12 phones in (6.6b) and (6.6c) are the same, and in almost the same order (by median). These 12 phones show a particularly large mismatch between the visual alignment and both the auditory and the audiovisual alignment, which suggests that for these phones [ə, ʔ, n, t, ɪ, d, g, l, ʁ, ç, h, iː] the training procedure in the visual-only case determined strongly different phone borders from the other two

cases. A possible explanation for this is that these phones do not produce prominent effects in the visual feature trajectories, which seems intuitive: since our visual features consist of tracked markers on the lips and face only (and not, e.g., motion features of the tongue or other intra-oral articulators), phones that do not have a strong effect on the movement of the lips and jaw are difficult to capture in the visual feature space. The consonants [ʔ, n, t, d, g, l, ʀ, ç, h] are all mainly defined by intra-oral articulation—in contrast to, e.g., the consonants [f, p, b, m, ʃ], which have a strong effect on lip motion and accordingly appear close to the top in Figure 6.6b and 6.6c. Likewise, it can be argued that the vowels [ə, ɪ, iː] exhibit rather indistinct lip motion, whereas diphthongs and rounded vowels can be expected to yield more characteristic trajectories.

## 6.5 Evaluation

In order to assess the quality of the various models and synchronization strategies described in Section 6.3, we have carried out a subjective evaluation experiment with 21 non-expert subjects (13 female, 15 male, aged 20 to 37, mean age 26.5) using a web-based experimental setup. For this experiment, 10 held-out test utterances from our recordings were synthesized using all methods and synchronization strategies and all of our four speakers. The evaluation consisted of an acoustic-only and an audiovisual part.[1]

### 6.5.1 Acoustic Evaluation

To investigate the effect on quality of the audio synthesis of the joint-audiovisual system by adding an additional visual stream, we have evaluated the different methods in a pair-wise comparison listening test. In each comparison, the listeners heard two audio samples from two different methods, but containing the same utterance from the same speaker. After hearing each sample as many times as they liked, they were asked to decide which of the two they preferred with respect to overall quality. No preference (a "tie") was also an option. Four methods for synthesizing audio were compared in this test: *audio*, which represents the regular audio-only system (and hence the synchronization strategies *unsync*, *uttlen-audio* and *durcopy-audio*), *audiovisual*, which represents the audio generated from the joint-audiovisually trained model, *durcopy-visual*, which represents audio synthesized with the visual duration model (used in the synchronization

[1] Example stimuli for all parts of the evaluation are available on `http://schabus.xyz/phd/audiovisual`.

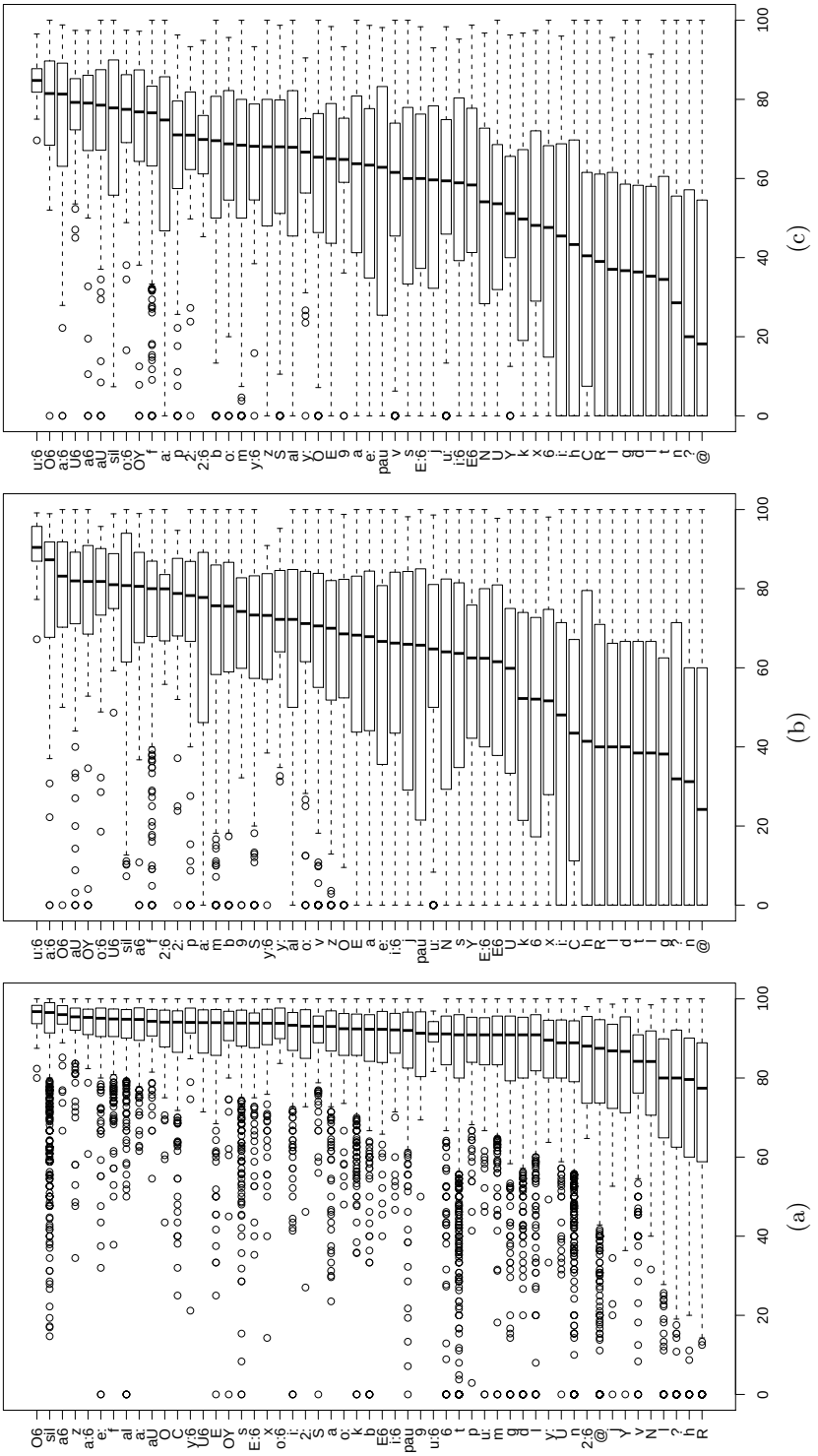



Figure 6.6: Boxplots for matching percentage per phone ($M_{phone}$). (a) Auditory and audiovisual. (b) Visual and audiovisual. (c) Auditory and visual.

Table 6.4: Evaluation results for the acoustic part.

| Compared Methods | | | wins | ties | sig. |
|---|---|---|---|---|---|
| recorded | : | audio | 76 : 3 | 1 | ∗ |
| recorded | : | audiovisual | 77 : 1 | 2 | ∗ |
| recorded | : | durcopy-visual | 79 : 1 | 0 | ∗ |
| audio | : | audiovisual | 19 : 11 | 50 | |
| audio | : | durcopy-visual | 44 : 6 | 30 | ∗ |
| audiovisual | : | durcopy-visual | 43 : 2 | 35 | ∗ |

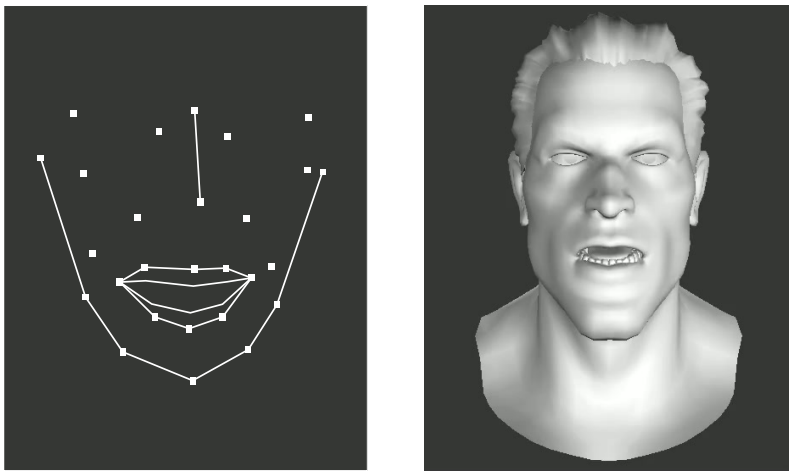
strategy of the same name), and original recorded speech (*recorded*).[2] All possible comparisons were heard twice by different listeners. The results are given in Table 6.4, where the "winning" scores and the number of ties are listed for each method pair. In the last column, the symbol "∗" indicates statistical significance of the score difference according to Bonferroni-corrected Pearson's $\chi^2$-tests of independence with $p < 0.01$.

Recorded audio was rated better than synthetic speech from any of the methods, and audio synthesized using the visual duration model (*durcopy-visual*) was rated worse than everything else. The small difference between *audio* and *audiovisual* (19 vs. 11) is not statistically significant ($p > 0.42$) and their similarity is also reflected in the large number of "ties" (50). We interpret these results to indicate that the additional visual stream in the joint audiovisual training has no significant effect (neither positive nor negative) on the quality of the generated acoustic speech signals.

### 6.5.2 Audiovisual Evaluation

In order to evaluate the audiovisual models and in particular the temporal alignment quality of the different synchronization strategies described in Section 6.3, we compared rendered videos consisting of synthesized facial motion and synthesized speech in the second part of the experiment. Similar to Bailly et al. (2002), to focus on evaluating the quality of the generated marker motion rather than the quality of the retargeting procedure or the appearance of the 3D head model, we have decided to present the raw synthesized marker motion to the subjects, i.e., renderings of the 27 points moving in 3D space, with some supporting lines added for orientation as shown in Figure 6.7a. The inner lip contours were added automatically based on a fixed distance between the outer lip markers and six corresponding points that define the inner lip. Even though this method does not

[2] We did not include the audio from the synchronization strategy *uttlen-visual*, because it is barely if at all distinguishable from *audio*, due to the small absolute values of $\rho$ in our experiments.



(a) (b)

Figure 6.7: Mode of speech motion presentation in the first and second (expert) evaluations. Example videos are available on `http://schabus.xyz/phd/audiovisual`. (a) Raw marker data. (b) Data-controlled 3D head.

necessarily produce all lip closures, it only generates correct lip closures. Note that in a setup with marker motion retargeting to a 3D head, these lines are not needed and all speech motion, including closures and lip compression, is computed based on the marker positions alone by the retargeting procedure.

In each pair-wise comparison in this part of the experiment, the subjects saw two videos from two different methods containing the same utterance from the same speaker. After watching each video as many times as they liked, they were asked to decide which of the two had better synchronization between acoustic speech and visible speech movement. No preference (a "tie") was also an option. As it was done in the LIPS 2008/2009 challenges (Theobald et al., 2008), we have chosen to ask specifically for synchronization quality, rather than testing more generally for intelligibility and naturalness.

In this test, we compared all synchronization strategies described in Section 6.3, as well as recorded speech and motion data, against each other. The results are given in Table 6.5, where the "winning" scores and the number of "ties" are listed for each method pair. In the last column, the symbol "$*$" indicates statistical significance of the score difference according to Bonferroni-corrected Pearson's $\chi^2$-tests of independence with $p < 0.05$.

The results in Table 6.5 confirm that recorded speech and recorded speech movements were perceived to be synchronized significantly better than any generated stimuli, and that *durcopy-visual* was perceived as having worse synchronization than the two *uttlen* methods. In particular, the *audiovisual*

Table 6.5: Evaluation results for the audiovisual part.

| Compared Methods | | | wins | ties | sig. |
|---|---|---|---|---|---|
| recorded | : | audiovisual | 32 : 5 | 3 | * |
| recorded | : | durcopy-audio | 25 : 7 | 8 | * |
| recorded | : | durcopy-visual | 32 : 6 | 2 | * |
| recorded | : | uttlen-audio | 24 : 9 | 7 | * |
| recorded | : | uttlen-visual | 26 : 8 | 6 | * |
| recorded | : | unsync | 25 : 11 | 4 | * |
| audiovisual | : | durcopy-audio | 9 : 17 | 14 | |
| audiovisual | : | durcopy-visual | 18 : 8 | 14 | |
| audiovisual | : | uttlen-audio | 10 : 10 | 20 | |
| audiovisual | : | uttlen-visual | 11 : 20 | 9 | |
| audiovisual | : | unsync | 9 : 14 | 17 | |
| durcopy-audio | : | durcopy-visual | 11 : 9 | 20 | |
| durcopy-audio | : | uttlen-audio | 6 : 11 | 23 | |
| durcopy-audio | : | uttlen-visual | 10 : 12 | 18 | |
| durcopy-audio | : | unsync | 12 : 12 | 16 | |
| durcopy-visual | : | uttlen-audio | 6 : 21 | 13 | * |
| durcopy-visual | : | uttlen-visual | 6 : 18 | 16 | * |
| durcopy-visual | : | unsync | 8 : 19 | 13 | |
| uttlen-audio | : | uttlen-visual | 8 : 14 | 18 | |
| uttlen-audio | : | unsync | 11 : 9 | 20 | |
| uttlen-visual | : | unsync | 9 : 9 | 22 | |

method only performed differently from the *recorded* condition but not from any other method. We expected the *audiovisual* method to be perceived as having the closest synchronization between the visual and the audio stream. However, there are several possible reasons for the absence of such a perceived synchronization:

- The utterances in the evaluation were short (4–7 words), randomly selected held-out test sentences from our recorded data. Longer sentences rich in phones that exhibit prominent lip motion (as identified in Section 6.4) might show stronger differences between the methods.
- The decision to present animated raw marker data rather than an animated 3D head model controlled by this data might have been a counter-productive one.
- The test subjects were non-experts recruited on the web, who might have only reported very obvious differences, resulting in "washed-out" results for the more subtle differences.

To further test the synchronization, an additional evaluation was carried out with subjects judging "challenging" utterances, which were longer (12–17 words), semantically unpredictable but syntactically correct utterances,

Table 6.6: Evaluation results using "challenging" utterances.

| Compared Methods | | experts wins | experts ties | experts sig. | non-experts wins | non-experts ties | non-experts sig. |
|---|---|---|---|---|---|---|---|
| audiovisual | : durcopy-audio | 15 : 5 | 5 | * | 25 : 24 | 16 | |
| audiovisual | : uttlen-audio | 17 : 3 | 5 | * | 34 : 22 | 9 | |
| audiovisual | : uttlen-visual | 17 : 4 | 4 | * | 31 : 23 | 11 | |
| durcopy-audio | : uttlen-audio | 10 : 8 | 7 | | 31 : 15 | 19 | * |
| durcopy-audio | : uttlen-visual | 10 : 8 | 7 | | 30 : 18 | 17 | |
| uttlen-audio | : uttlen-visual | 5 : 6 | 14 | | 18 : 25 | 22 | |

rich in audiovisual "landmarks", synthesized following the four synchronization strategies *audiovisual*, *uttlen-audio*, *uttlen-visual* and *durcopy-audio*. We do not have recordings of these utterances and we excluded the *durcopy-visual* strategy because of its bad performance in the first evaluation. We also excluded *unsync* because of the strong similarity of this method to the two *uttlen* methods. We applied the synthesized marker motion to a 3D head model via retargeting and created rendered animation sequences from these (see Figure 6.7b for an example frame). 13 non-expert subjects and 5 expert subjects (speech technology, phonetics) took part in this evaluation (9 female, 9 male, aged 22 to 58, mean age 33.9). Otherwise the experimental setup was identical to the first evaluation. The results are given in Table 6.6.

For these "challenging" utterances, the experts perceived the *audiovisual* method to produce significantly better speech/motion synchronization than the other methods, which show no significant difference among each other. For the non-expert subjects, on the other hand, the only significant difference is between *ducropy-audio* and *uttlen-audio*. This suggests that the *audiovisual* method produces improved synchronization, but some subtle differences are not consciously perceived by the non-expert subjects, although a clear trend in favor of the *audiovisual* method is also visible for the non-experts.

Figure 6.8 shows excerpts of synthesized trajectories for one of the "challenging" utterances. The top part of the figure illustrates the *uttlen-visual* strategy. Although identical total utterance duration is ensured, the two duration models generate different phone durations within the utterance, resulting in a clear misalignment of some feature "landmarks", as indicated in the figure by dashed magenta lines. The middle part of the figure illustrates the joint audiovisual strategy. The single audiovisual duration model provides better alignment of the same feature "landmarks". It is quite obvious that this causes a perceptible improvement over the *uttlen-visual* method. The bottom part illustrates the *durcopy-audio* method. Overwriting the visual

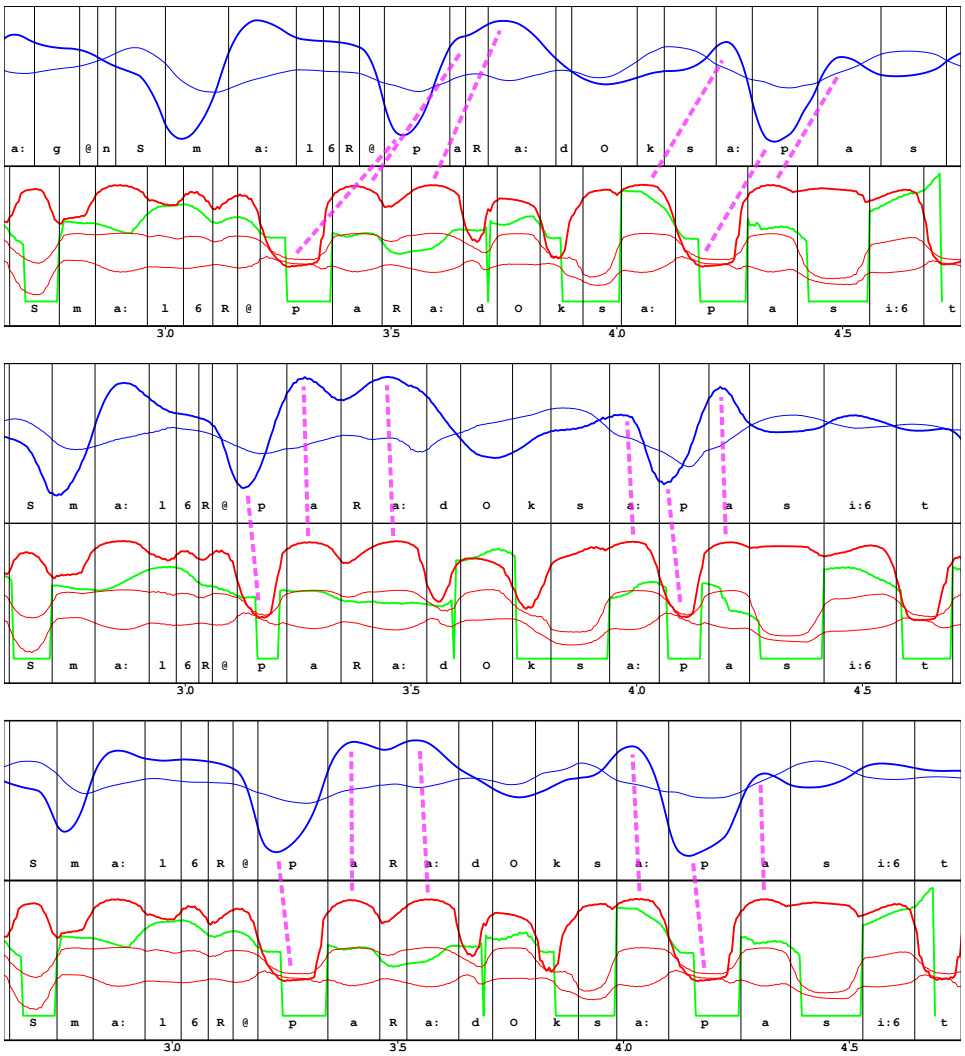



Figure 6.8: Excerpts of synthesized audiovisual trajectories for one of the "challenging" utterances from different synthesis strategies: uttlen-visual (top), audiovisual (middle) and durcopy-audio (bottom). The plots show the Euclidean distance between the central upper lip and central lower lip markers (thick blue line), the Euclidean distance between the left and right mouth corner markers (thin blue line), the first three MFCCs (red, with decreasing thickness) and F0 (green). The different features have been re-scaled to fit into the same vertical range. Some feature landmark correspondences are indicated by magenta dashed lines. The sentence is "Selbst wenn der Nottaste beim Nachsagen schmalere Paradoxa passierten fragt der Bahnschaffner fahrerder Wagen weiter" (even if narrower paradoxes happened to the emergency button while repeating, the train conductor of moving cars continues to ask).

duration model with the audio one guarantees alignment of the phone borders, resulting in good alignment also of the feature "landmarks". However, forcing the visual system to use predefined durations can result in artificial contraction or stretching of phones, leading to unnaturally fast or slow movement, as visible in the stretched [p] phone between second 4 and 4.5. As the expert evaluation has shown, this leads to a perceptible inferiority of this synchronization strategy to the *audiovisual* method.

## 6.6 Conclusion

In this chapter we showed that joint audiovisual speech synthesis improves the quality of the visual speech compared to other synchronization approaches. In our first evaluation we saw no differences between audiovisual modeling and other synchronization approaches, except for the recorded data which was always better than the models. Concerning acoustic synthesis quality, all models except *audiovisual* performed worse than acoustic modeling only.

During an additional evaluation with visually challenging utterances, the audiovisual model performed significantly better than other synchronization approaches when judged by expert listeners. In addition, the analysis of the state-alignments, produced by the different models, showed objective differences in audiovisual alignment between the proposed approaches. In summary the proposed integrated speaker-dependent audiovisual approach allows for joint modeling of visual and acoustic signals while maintaining high-quality acoustic synthesis results with improved audiovisual synchronization over other methods.

# Chapter 7

# Speaker-Adaptive Audiovisual Speech Synthesis

In this chapter, which is closely related to the earlier publication of Schabus et al. (2012b), we apply speaker-adaptive and speaker-dependent training of hidden Markov models (HMMs) to visual speech synthesis. In speaker-dependent training we use data from one speaker to train a visual and acoustic HMM. In speaker-adaptive training, first a visual background model (average voice) from multiple speakers is trained. This background model is then adapted to a new target speaker using (a small amount of) data from the target speaker. This concept has been successfully applied to acoustic speech synthesis. This chapter demonstrates how model adaption is applied to the visual domain, synthesizing animations of talking faces. A perceptive evaluation is performed, showing that speaker-adaptive modeling outperforms speaker-dependent models for small amounts of training / adaptation data.

## 7.1 Introduction

The goal of audiovisual text-to-speech synthesis is to generate both an acoustic speech signal as well as a matching animation sequence of a talking face, given some unseen text as input. Most commonly, acoustic and visual synthesis are addressed separately, and although we have addressed joint audiovisual modeling in Chapter 6, we follow the separated approach in this chapter.

In the preceding chapters, we have already shown that the popular acoustic

HMM-based speech synthesis system HTS can be extended to the visual domain for training and synthesis of 3D facial motion control parameters. However, as with all HMM-based approaches, large amounts of training data are required to build high quality systems and recording large amounts of visual data is even more costly than recording audio data. To address this shortcoming for speakers where limited amounts of data are available, a very successful speaker-adaptive approach has been developed (Yamagishi and Kobayashi, 2007; Yamagishi et al., 2009a) for acoustic HMM-based speech synthesis, as presented briefly in Chapter 3. A (possibly large) speech database containing multiple speakers is used to train an average voice, where a speaker-adaptive training (SAT) algorithm provides speaker normalization. Then, a voice for a new target speaker can be created by transforming the models of the average voice via speaker adaptation, using (a possibly small amount) of speech data from the target speaker. This allows the creation of many speakers' synthetic voices without requiring large amounts of speech data from each of them. It can be shown that synthetic speech from voice models created in this way is perceived as more natural sounding than synthetic speech from speaker-dependent voice models using the same (target speaker) data (Yamagishi and Kobayashi, 2007). This holds especially for the case where this data is of small amount. The goal of this chapter is to demonstrate how this speaker-adaptive training approach can be applied to visual speech synthesis.

## 7.2 Adaptive visual speech synthesis system

The CSTR/EMIME training scripts for the HTS system that were extended to visual and audiovisual synthesis (cf. Chapter 4) are also provided in a speaker-adaptive version for training average voices across multiple speakers and in a target speaker adaptation version. For the investigations in this chapter, these were modified for visual speech synthesis, in a similar way as the speaker-dependent versions. The resulting speaker-adaptive visual modeling framework is illustrated in Figure 7.1 (cf. Figure 3.8 from Chapter 3). It consists of a training, adaptation, and synthesis module. Context-dependent, left-to-right, hidden semi-Markov models (HSMMs) are trained on multi-speaker visual databases in order to simultaneously model the visual features, as well as duration. We use SAT based on CMLLR for the training of the average visual models and for adaptation also CMLLR (Yamagishi and Kobayashi, 2007; Yamagishi et al., 2009a). The visual feature extraction is applied to a multi-speaker database before training, and to a possibly different single speaker database before adaptation. In the synthesis step, visual parameters are generated from the adapted models.

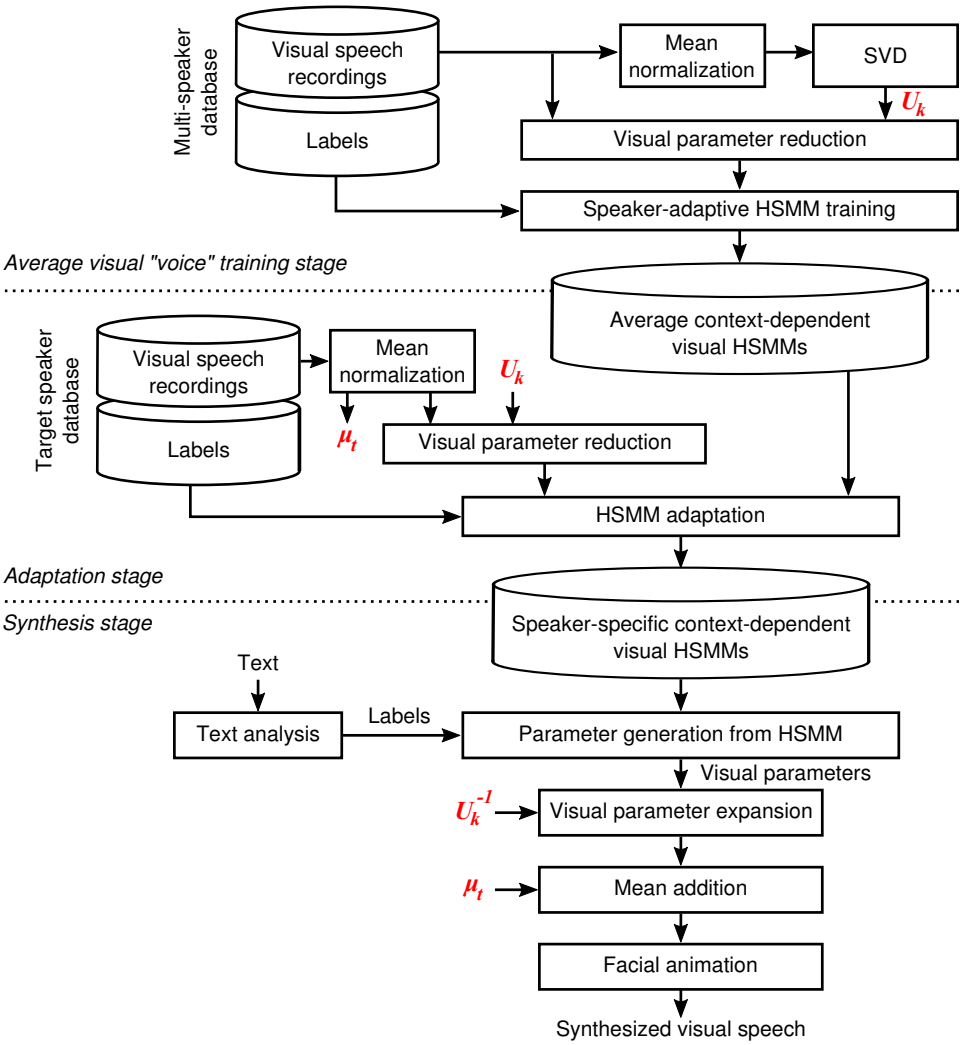



Figure 7.1: Overview of the speaker-adaptive visual modeling framework.

As exemplary data, the FMSC corpus was used, i.e., audiovisual speech recordings of one female and two male speakers of Standard Austrian German (cf. Chapter 5). The recordings consisted of 223 utterances (roughly 11 minutes) per speaker. Figure 7.2 shows example frames.

As described in Chapter 4, the visual feature extraction we use for the training of the average visual voice first applies mean normalization and SVD to derive a matrix $\boldsymbol{U}_k$ that is used to project the recorded marker data to a lower $k$-dimensional space. In the adaptation step we also perform mean normalization using the speaker mean $\boldsymbol{\mu}_s$ and then use $\boldsymbol{U}_k$ from average voice training to reduce the visual adaptation features. In visual synthesis, the generated features are projected back to the full feature space using $\boldsymbol{U}_k^{-1}$,

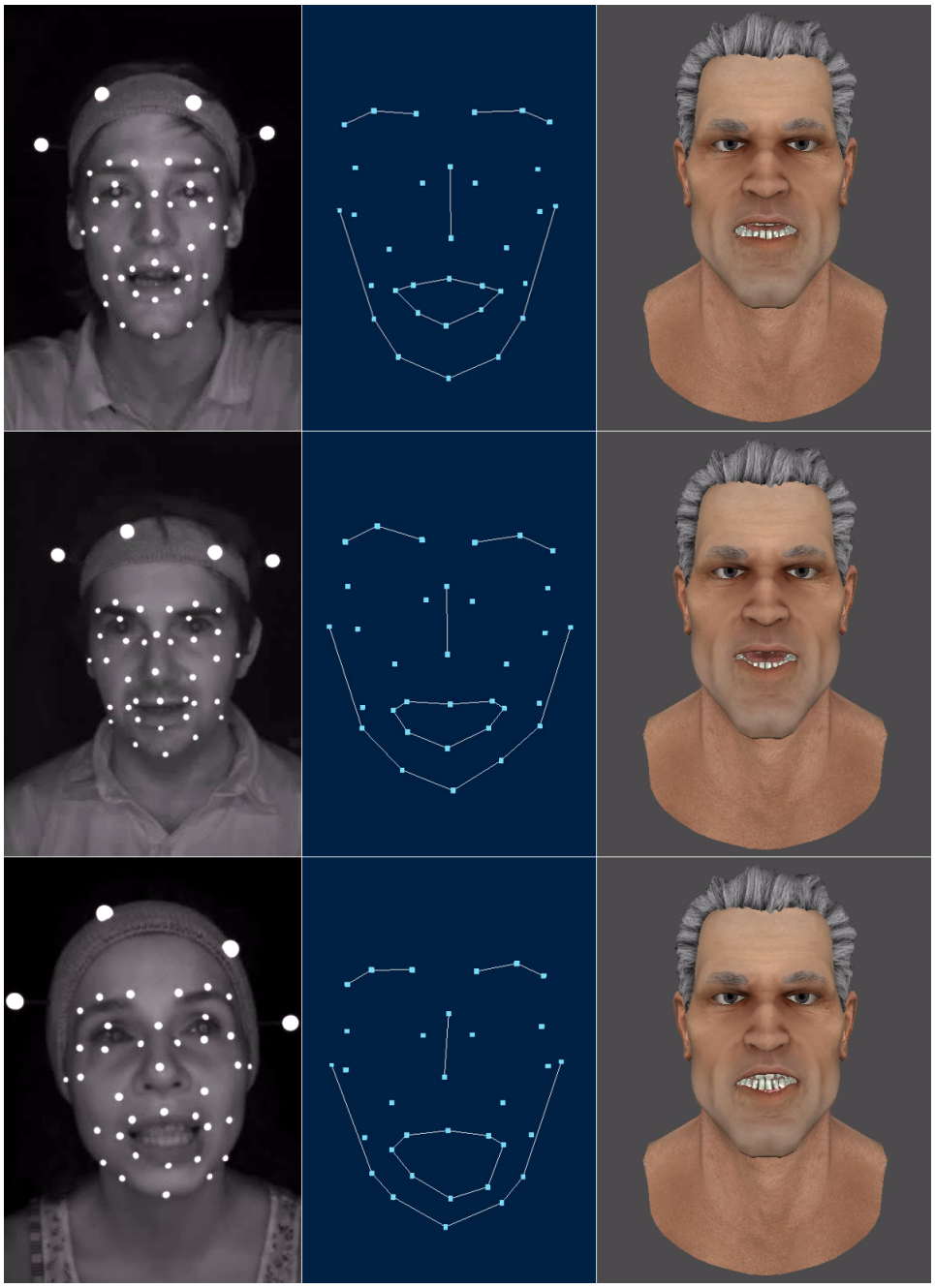


Figure 7.2: Still images from the recording session (left), the corresponding 3D marker data (middle) and the resulting pose of the virtual head with this data applied (right). See also videos at http://schabus.xyz/phd/adaptation.

and the speaker mean $\boldsymbol{\mu}_s$ is added. The resulting visual features are used to animate a talking head.

We would like to emphasize that the feature projection matrix $\boldsymbol{U}_k$ is the same in the training, adaptation and synthesis steps, and that it is determined via SVD without using data from the target speaker, i.e., in the entire process there is only one SVD calculation, namely across all speakers that contribute to the average voice. The speaker means, on the other hand, are subtracted per speaker before SVD and projection in the training part, and also before projection in the adaptation part.

In speaker-dependent modeling, the training data comes from one speaker $s$, $\boldsymbol{U}_k$ and $\boldsymbol{\mu}_s$ are determined on that speaker's data and the whole adaptation step is missing.

## 7.3 Evaluation

To evaluate our system, 10 held-out test utterances where visually synthesized. In order to allow for direct comparison of recorded data to synthesized utterances, the true phone durations from the recorded data were employed instead of generated durations from the trained duration models. This results in all stimuli from the same speaker and utterance to be of equal length on a phone-by-phone basis. In terms of the synchronization strategies discussed in Chapter 6, the setup here is similar to the *durcopy-audio* strategy (cf. Section 6.3), except that the phone durations are not predicted from the audio duration model but taken from the forced alignment results on the recorded data.

We compare the recorded visual data (which we will refer to as *recorded*) to four training strategies: 1) The speaker-adaptive method we presented in the previous section, where an average voice is trained on the data of two speakers (212 utterances each), which is then adapted to the third speaker using also 212 utterances (*adapted*). 2) A corresponding speaker-dependent model, trained on the target speaker's 212 utterances (*sd*). 3) An adapted model with a small amount of adaptation data; here, the average voice is the same as in *adapted*, but for adaptation we use the smallest set of utterances that contains each phone at least three times (19 utterances) (*adapt small*). 4) A speaker-dependent model trained on the same small amount of data (*sd small*).

Similar to our objective reconstruction error calculations during the analysis of the PCA projections (Chapter 4), we have computed objective errors by calculating the frame-wise deviations of marker positions between recorded and synthesized sequences. Figure 7.3 shows the resulting RMSE, calculated

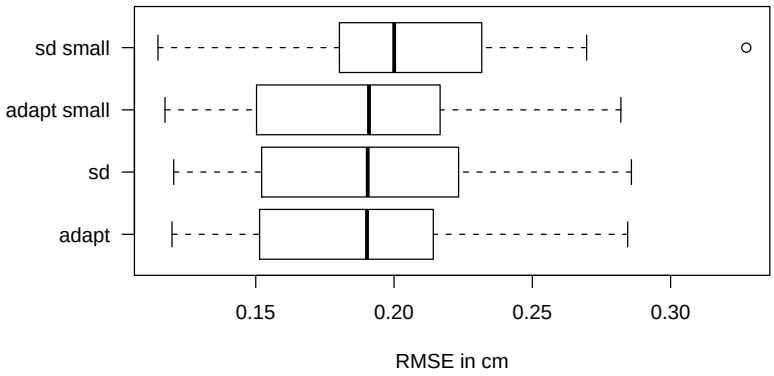



Figure 7.3: Box plots of the root mean squared differences between synthesized and recorded marker positions.

across all frames of each utterance. Since we have 10 test utterances and three speakers, each box plot contains 30 RMSE values. Unfortunately, these objective results are not very informative. If anything, we can observe that the RMSE for *sd small* is slightly larger than for the other methods. This is mainly due to temporal misalignment: although we force the parameter generation to produce the same phone durations as the ones in the recorded data, slight temporal shifts of the valleys and peaks of a trajectory within a phone can cause a large error even though the movement of the corresponding maker is "correct". Objective evaluation of synthesized marker motion by comparison to recorded data is therefore not straightforward.

Therefore, we have conducted a perceptive experiment with 28 test subjects (11 female, 17 male, aged 15 to 49, mean age 27.5). Each subject saw 45 pairs of videos showing a virtual head driven by two different models (*recorded*, *sd*, *sd small*, *adapted*, *adapted small*), where all possible combinations of methods, speakers and utterances were distributed among the subjects such that each subject saw each of the ten method combinations, as well as each speaker-utterance at least once. To each video we have added a synthetic speech sample generated from models that we trained on the corresponding speaker's acoustic data from our synchronous corpus. As for the visual synthesis, we have provided the phone borders from the recordings rather than using the duration model.[1]

For each video pair, the subjects selected whether they preferred the first or the second video, or they thought they were of equal quality. The results are given in Table 7.1, where we have counted the number of "won" comparisons and the number of "ties" for each method pair. To assess the statistical significance of these preference scores, we have computed Bonferroni-corrected Pearson's $\chi^2$-tests of independence with $p < 0.01$ for each method pair. The

[1] See example stimuli at `http://schabus.xyz/phd/adaptation`.

Table 7.1: Pair-wise comparison scores.

| Compared methods | | | wins | ties | sig. |
|---|---|---|---|---|---|
| recorded | : | sd | 74 : 33 | 20 | ∗ |
| recorded | : | sd small | 95 : 25 | 10 | ∗ |
| recorded | : | adapt | 95 : 20 | 10 | ∗ |
| recorded | : | adapt small | 86 : 22 | 10 | ∗ |
| sd | : | sd small | 64 : 36 | 22 | ∗ |
| sd | : | adapt | 54 : 39 | 28 | |
| sd | : | adapt small | 56 : 37 | 39 | |
| sd small | : | adapt | 56 : 34 | 35 | |
| sd small | : | adapt small | 31 : 57 | 37 | ∗ |
| adapt | : | adapt small | 27 : 35 | 73 | |

results are given in the last column of Table 7.1, where the symbol "∗" indicates a statistically significant influence of the methods on the preference scores.

The animations that replay the recorded data are preferred significantly more times over all the synthesis methods. Furthermore, within the speaker-dependent methods *sd* and *sd small*, the reduction in training data results in a significant difference between the two. The result between *sd* and *adapt* is not significant, but shows a trend towards the speaker-dependent model. However, *adapt small* is preferred over *sd small*, and the difference is statistically significant.

Summarizing, this chapter demonstrated how to apply average voice training and speaker adaptation to visual speech synthesis. This is useful when creating new systems for speakers where very few training utterances are available. In addition, with limited amount of training data the speaker adaptive approach outperforms speaker-dependent training. However, several additional experiments need to be conducted in future work. In particular, speaker similarity, a measure of how close synthesized data mimics specific speaker characteristics, needs to be investigated.

# Chapter 8

# Conclusion

This last chapter concludes the dissertation, by summarizing the preceding chapters and the respective scientific findings that were reported there, and by providing an outlook to possible future research, where questions could be addressed that have been left unanswered here or are otherwise related to this body of work.

## 8.1 Summary

This dissertation presented our investigations of audiovisual speech synthesis using hidden Markov models, i.e., the problem of generating audible speech and matching facial movement parameters for a 3D head model for any given textual input, using a statistical machine learning approach. While the application of the HMM-framework for (acoustic) speech synthesis to (audio-)visual signals is not novel in itself, several particular investigations that have been presented here do make a contribution to the advancement of the field, and those investigations have accordingly been published at relevant international journals and conferences.

Perhaps the most significant contribution—and certainly the most important of the published papers—is the study on joint audiovisual modeling presented in Chapter 6 (and previously in Schabus et al., 2014a). Most approaches to audiovisual speech synthesis that have been published to date treat the acoustic part as given and fixed and instead focus on the visual parts only. In order to synchronize the generated visual signal to the acoustic one, the phone borders as predicted by the acoustic system are passed on to the visual system. However, the detailed objective analysis in Chapter 6 has shown that in the HMM framework, the phonetic borders resulting from visual-only modeling are much less reliable than those resulting from

acoustic-only or joint audiovisual modeling. This suggests that the split into an audio-only and visual-only part may result in both weaker synchronization and degraded quality of visual modeling. In line with these findings, the subjective evaluation of the proposed truly joint audiovisual modeling approach showed a noticeable improvement over separate modeling.

Chapter 7 (and previously Schabus et al., 2012b) presented a study on speaker-adaptive visual speech synthesis. The concept of adaptation, where first an average voice of multiple speakers is trained, which is then adapted towards a specific target speaker, has been one of the key factors in the success of the HMM-based speech synthesis framework. In spite of this, adaptation had not been previously explored for the visual speech domain. The experiments described in Chapter 7 showed that this concept is indeed applicable also to the visual domain, and that the adaptive method does outperform the "classical" speaker-dependent method, at least when the amount of data available from the target speaker is small.

As a pre-requisite for visual speaker-adaptive training, a suitable representation of the visual signal is required. The feature extraction part of Chapter 4 (and previously Schabus et al., 2012a and Schabus et al., 2013) addressed this issue. Principal Component Analysis (PCA) can provide dimensionality reduction and component de-correlation, resulting in feature vectors that are well-suited for statistical modeling, because diagonal instead of full covariance matrices may be used. The objective and subjective evaluation experiments described in Chapter 4 furthermore showed that a common subspace for multiple speakers can be found via PCA and that such a sub-space is also general enough to contain new (target) speakers, as long as the number of dimensions is not reduced too aggressively; hence this method of visual feature extraction is suitable for adaptive visual speech modeling.

In order to be able to carry out the aforementioned experiments, a pipeline for recording, feature extraction, model training, synthesis and animation rendering had to be developed, as described in Chapter 4 (and, to some degree, previously in Schabus et al., 2011, Schabus et al., 2012b and Schabus et al., 2014a). For the core part—model training and synthesis—the well-known acoustic speech synthesis framework HTS was extended to additionally utilize visual speech features. Furthermore, several original software components were developed, for example for visual feature extraction, semi-automated data cutting and synchronization, point cloud visualization, batch retargeting and rendering, objective evaluations via distance measure calculations, and subjective evaluations via web-based user trials.

Finally, adequate data is required for experiments in data-driven speech signal processing. In contrast to acoustic speech, corpora of speech motion tracking data are not readily available from the research community. Therefore, three corpora of synchronous speech recordings with 3D facial

motion tracking data have been created within this dissertation project, as described in Chapter 5 (and previously in Schabus et al., 2012a, Schabus et al., 2014a and Schabus et al., 2014b). Recordings in Standard Austrian German from eleven speakers, Austrian dialectal recordings from eight speakers, and recordings of varying speaking rate with additional tongue motion tracking from one speaker were recorded, pre-processed, labeled and manually refined. Most of this data is already available on the Internet for research purposes, and an additional release of the remaining data is planned for the future.

In addition to presenting the conducted research, this dissertation also attempted to give a broad overview over its background and related work, covering other approaches to (audio-)visual speech synthesis in Chapter 2 and the HMM framework for (acoustic) speech synthesis in Chapter 3.

## 8.2 Outlook

Several open questions can be identified that seem to be relevant to the field of audiovisual speech synthesis from the perspective of the end of this multi-year research project.

The adaptation paradigm has not yet been exhaustively investigated for visual and audiovisual speech synthesis. Chapter 7 has addressed this topic and shown the principal applicability of the adaptive approach to the visual domain. However, larger experiments seem to be necessary, using data from many more speakers and conducting more extensive evaluations, with the goal to show much clearer benefits of the adaptive approach than Chapter 7 was able to deliver. Such experiments should also include studies on the achieved target speaker similarity of the generated facial motion, using objective and subjective measures. Due to the large influence of the 3D head and the retargeting function on the perception of the end result, this may turn out to be quite difficult. Showing the need for speaker-specific facial motion is however crucial for "justifying" adaptation, because otherwise a single high-quality face motion model may be used for all speakers. With increasing fidelity of facial motion capturing, especially marker-less systems that directly capture a facial mesh of high resolution, speaker-specific motion models can be expected to become more important than they are for the kind of data used in this dissertation. In addition to average voices and adaptation, audiovisual data from multiple speakers allows experiments on mixed-speaker setups, speaker swapping, speaker interpolation etc. in the visual and audiovisual domain.

Different performance capturing methods and other data sources are generally an interesting topic from a synthesis point of view. The same HMM-

based trajectory modeling methodology can be applied to any parametrization of facial motion over time, provided that it exhibits sufficient temporal smoothness and hence predictability within short intervals. For high-density facial surface recordings, this might require a dimensionality reduction step such as PCA, as described in Chapter 4. Such recording collections may also help to automate the creation of (marker to full mesh) retargeting functions, by placing "virtual" markers on the recorded high-resolution face and learning their influence on the entire face using some sort of regression. Furthermore, large quantities of high-quality facial parameter data exist in animated films and computer games. This kind of data is typically recorded and then extensively improved by animation experts, if not entirely created manually, and thus would make an interesting body of training data, if it were available. Independent of the data source, automatic retargeting generally remains an interesting open challenge for the field of 3D audiovisual speech synthesis, especially concerning lip closures and non-rigid lip deformations.

Chapter 6 has touched on the topic of visemes/phoneme equivalence classes. This topic, or more generally speaking, the choice of the optimal phone inventory for audiovisual speech modeling, remains an open problem, again requiring objective and especially subjective evaluation experiments.

As mentioned in Chapter 5, audiovisual recordings of Austrian dialect speakers have been created during this project. While it is straightforward to create dialectal audiovisual synthesizers from this kind of data, other interesting problems can be investigated, like dialect/standard interpolation, cross-mapping of dialects and speakers, and dialect adaptation, all of which are being investigated for acoustic speech synthesis but not yet for the visual domain. Before turning to the modeling part of these problems, it should be verified that the choice of language variety (or even the choice of language) within the same speaker does actually have an objectively and/or subjectively detectable influence on the facial motion parameters.

A special data corpus that includes electromagnetic tongue motion recordings in addition to optical face marker data was presented in Chapter 5. Some preliminary data analysis was carried out on this data, but there are several additional interesting ways in which this corpus could be used. For example synthesis including tongue motion, which should be straightforward to realize using the same methods as presented for face motion in this dissertation. However, a 3D head with retargeting also defined for the tongue would be required. As another example, modification of acoustic parameters by visual control or combined articulatory-visual control could be realized, similar to existing articulatory-to-acoustic control models.

In *acoustic* speech synthesis research, there is a generally accepted consensus that the quality of generated speech cannot be assessed using objective

measures alone, but that subjective listening experiments are required to obtain meaningful evaluation results. Throughout this project, it has become evident that the same is true also for visual speech signals. Furthermore, it has turned out that subjective evaluation of facial motion seems to be a very difficult task. Many factors, like the appearance of the 3D head and the quality of the retargeting function play an important role and may mask small differences in the generated parameter trajectories. On the other hand, several experiments showed that presenting the raw marker motion as point cloud animations does also not seem to be a good alternative, perhaps because the resulting stimuli are inherently unnatural and quite uniform. It remains a challenge to define how to best evaluate synthesized visual speech, especially if we think about comparing different systems/approaches to each other.

For animated films and computer games, merely producing "correct" speech and speech motion is not sufficient to convey a particular story. Emotional and conversational speech need to be produced, accompanied by non-verbal facial and vocal behavior like facial expressions, laughter, yawning, sneezing, screaming, etc. This was not addressed in this dissertation, but it is a very important and at the same time extremely difficult challenge for the field.

# Bibliography

P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux (2002). "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images". In: *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553 (cit. on p. 40).

P. Badin, P. Borel, G. Bailly, L. Revéret, M. Baciu, and C. Segebarth (2000). "Towards an Audiovisual Virtual Talking Head: 3D Articulatory Modeling of Tongue, Lips and Face Based on MRI and Video Images". In: *Proceedings of the 5th Speech Production Seminar*. Kloster Seeon, Germany, pp. 261–264 (cit. on pp. 40, 41).

L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek, B. Lewis, and R. Mercer (1980). "Further results on the recognition of a continuously read natural corpus". In: *Proceedings of the 1980 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 5. Denver, CO, USA, pp. 872–875 (cit. on p. 63).

L. Bahl, P. V. de Soutza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny (1991). "Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees". In: *Proceedings of the Workshop on Speech and Natural Language*. Pacific Grove, CA, USA, pp. 264–269 (cit. on p. 63).

G. Bailly, M. Bérar, F. Elisei, and M. Odisio (2003). "Audiovisual Speech Synthesis". In: *International Journal of Speech Technology*, vol. 6, no. 4, pp. 331–346 (cit. on p. 31).

G. Bailly, G. Gibert, and M. Odisio (2002). "Evaluation of movement generation systems using the point-light technique". In: *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*. Santa Monica, CA, USA, pp. 27–30 (cit. on p. 128).

G. Bailly, O. Govokhina, G. Breton, F. Elisei, and C. Savariaux (2008). "A Trainable Trajectory Formation Model TD-HMM Parameterized for the LIPS 2008 Challenge". In: *Proceedings of the 9th Annual Conference of*

*the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, pp. 2318–2321 (cit. on p. 43).

G. Bailly, O. Govokhina, F. Elisei, and G. Breton (2009). "Lip-Synching Using Speaker-Specific Articulation, Shape and Appearance Models". In: *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, no. 769494, pp. 1–11 (cit. on pp. 43, 44, 47, 114, 122).

G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, eds. (2012). *Audiovisual Speech Processing*. Cambridge, UK: Cambridge University Press (cit. on p. 31).

J. A. Baker (1975). "The DRAGON system–An overview". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 24–29 (cit. on p. 27).

L. E. Baum, T. Petrie, G. Soules, and N. Weiss (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171 (cit. on p. 56).

M. Baum, G. Erbach, and G. Kubin (2000). "SpeechDat-AT: A telephone speech database for Austrian German". In: *Proceedings of the LREC Workshop Very Large Telephone Databases (XL-DB)*. Athens, Greece, pp. 51–56 (cit. on p. 19).

T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross (2011). "High-quality Passive Facial Performance Capture Using Anchor Frames". In: *Proceedings of the 38th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. Vancouver, BC, Canada, 75:1–75:10 (cit. on pp. 45, 46).

L. E. Bernstein (2012). "Visual Speech Perception". In: *Audiovisual Speech Processing*. Ed. by G. Bailly, P. Perrier, and E. Vatikiotis-Bateson. Cambridge University Press, pp. 21–39 (cit. on p. 118).

J. Beskow (1995). "Rule-Based Visual Speech Synthesis". In: *Proceedings of the 4th European Conference on Speech Communicatino and Technology*. Madrid, Spain, pp. 199–302 (cit. on pp. 37, 39).

J. Beskow (1997). "Animation of Talking Agents". In: *Proceedings of the 1st International Conference on Auditory-Visual Speech Processing (AVSP)*. Rhodes, Greece, pp. 149–152 (cit. on p. 39).

J. Beskow (2003). "Talking Heads - Models and Applications for Multimodal Speech Synthesis". PhD thesis. Stockholm, Sweden: KTH Stockholm (cit. on pp. 31, 37, 38, 40, 103, 110, 111).

J. Beskow (2004). "Trainable articulatory control models for visual speech synthesis". In: *International Journal of Speech Technology*, vol. 7, no. 4, pp. 335–349 (cit. on pp. 37, 39).

J. Beskow, O. Engwall, and B. Granström (2003). "Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements". In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*. Barcelona, Spain, pp. 431–434 (cit. on p. 39).

J. Beskow and M. Nordenberg (2005). "Data-Driven Synthesis of Expressive Visual Speech Using an MPEG-4 Talking Head". In: *Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH/INTERSPEECH)*. Lisbon, Portugal, pp. 793–796 (cit. on p. 40).

M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal (1999). "The AT&T next-gen TTS system". In: *Proceedings of the Joint meeting of ASA, EAA, and DAGA*. Berlin, Germany, pp. 18–24 (cit. on p. 26).

A. P. Breen and P. Jackson (1998). "A Phonologically Motivated Method of Selecting Non-Uniform Units". In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, pp. 2735–2738 (cit. on p. 26).

C. Bregler, M. Covell, and M. Slaney (1997). "Video Rewrite: driving visual speech with audio". In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. Los Angeles, CA, USA, pp. 353–360 (cit. on pp. 31, 32).

F. Brugnara, D. Falavigna, and M. Omologo (1993). "Automatic segmentation and labeling of speech based on Hidden Markov Models". In: *Speech Communication*, vol. 12, no. 4, pp. 357–370 (cit. on p. 53).

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan (2008). "IEMOCAP: interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359 (cit. on p. 100).

T. Chen (2001). "Audiovisual speech processing". In: *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21 (cit. on p. 118).

M. M. Cohen, J. Beskow, and D. W. Massaro (1998). "Recent Developments In Facial Animation: An Inside View". In: *Proceedings of the 2nd International Conference on Auditory-Visual Speech Processing (AVSP)*. Terrigal, Australia, pp. 201–206 (cit. on pp. 37–39).

M. M. Cohen and D. W. Massaro (1993). "Modeling Coarticulation in Synthetic Visual Speech". In: *Models and Techniques in Computer Anima-*

*tion*. Ed. by N. Magnenat-Thalmann and D. Thalmann. Springer-Verlag, pp. 139–156 (cit. on pp. 37, 38).

M. M. Cohen, D. W. Massaro, and R. Clark (2002). "Training a talking head". In: *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)*. Pittsburgh, PA, USA, pp. 499–504 (cit. on pp. 38, 39).

C. H. Coker, P. B. Denes, and E. N. Pinson (1963). *Speech Synthesis*. Baltimore, MD, USA: Waverly Press, Inc. (cit. on pp. 29, 30).

G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile (2000). "Segment Selection in the L&H RealSpeak Laboratory TTS System". In: *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. Beijing, China, pp. 395–398 (cit. on p. 26).

E. Cosatto (2002). "Sample-Based Talking-Head Synthesis". PhD thesis. Lausanne, Switzerland: Swiss Federal Institute of Technology (cit. on p. 32).

E. Cosatto and H. P. Graf (2000). "Photo-realistic talking-heads from image samples". In: *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163 (cit. on p. 32).

E. Cosatto, G. Potamianos, and H. P. Graf (2000). "Audio-visual unit selection for the synthesis of photo-realistic talking-heads". In: *IEEE International Conference on Multimedia and Expo (ICME)*. New York, NY, USA, pp. 619–622 (cit. on p. 32).

S. B. Davis and P. Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366 (cit. on pp. 34, 50).

A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38 (cit. on p. 56).

Z. Deng and U. Neumann, eds. (2008). *Data-Driven 3D Facial Animation*. London, UK: Springer (cit. on p. 31).

V. V. Digalakis, D. Rtischev, and L. G. Neumeyer (1995). "Speaker adaptation using constrained estimation of Gaussian mixtures". In: *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366 (cit. on p. 71).

P. Domingos (1999). "The Role of Occam's Razor in Knowledge Discovery". In: *Data Mining and Knowledge Discovery*, vol. 3, no. 4, pp. 409–425 (cit. on p. 66).

R. E. Donovan (1996). "Trainable Speech Synthesis". PhD thesis. Cambridge, UK: University of Cambridge (cit. on p. 59).

R. E. Donovan and E. M. Eide (1998). "The IBM Trainable Speech Synthesis System". In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, pp. 1703–1706 (cit. on p. 26).

R. E. Donovan and P. C. Woodland (1995a). "Automatic speech synthesiser parameter estimation using HMMs". In: *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. Detroit, MI, USA, pp. 640–643 (cit. on pp. 59, 60).

R. E. Donovan and P. C. Woodland (1995b). "Improvements in an HMM-Based Speech Synthesiser". In: *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH)*. Madrid, Spain, pp. 573–576 (cit. on p. 59).

T. Dutoit (1997). *An introduction to text-to-speech synthesis*. Norwell, MA, USA: Kluwer Academic Publishers (cit. on pp. 24, 25).

F. Elisei, M. Odisio, G. Bailly, and P. Badin (2001). "Creating and Controlling Video-Realistic Talking Heads". In: *Proceedings of the 4th International Conference on Auditory-Visual Speech Processing (AVSP)*. Aalborg, Denmark, pp. 90–97 (cit. on pp. 41, 42, 45).

T. Ezzat, G. Geiger, and T. Poggio (2002). "Trainable videorealistic speech animation". In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. San Antonio, TX, USA, pp. 388–398 (cit. on pp. 32, 33).

C. G. Fisher (1968). "Confusions Among Visually Perceived Consonants". In: *Journal of Speech, Language, and Hearing Research*, vol. 11, pp. 796–804 (cit. on p. 118).

J. E. Flege (1988). "Effects of speaking rate on tongue position and velocity of movement in vowel production". In: *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 901–916 (cit. on p. 110).

O. Fujimura (1968). "An Approximation to Voice Aperiodicity". In: *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 1, pp. 68–72 (cit. on p. 52).

T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai (1992). "An adaptive algorithm for mel-cepstral analysis of speech". In: *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. San Francisco, CA, USA, pp. 137–140 (cit. on p. 51).

S. Furui (1986). “Speaker-independent isolated word recognition using dynamic features of speech spectrum”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59 (cit. on p. 59).

S. Furui (2001). *Digital Speech Processing, Synthesis, and Recognition (Second Edition, Revised and Expanded)*. Marcel Dekker, Inc. (cit. on pp. 24, 25).

J.-L. Gauvain and C.-H. Lee (1994). “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”. In: *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298 (cit. on p. 71).

O. Govokhina, G. Bailly, and G. Breton (2007). “Learning optimal audiovisual phasing for a HMM-based control model for facial animation”. In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW)*. Bonn, Germany, pp. 1–4 (cit. on pp. 44, 47, 114, 122).

O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw (2006). “TDA: A New Trainable Trajectory Formation System for Facial Animation”. In: *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH)*. Pittsburgh, PA, USA, pp. 2474–2477 (cit. on p. 43).

T. Hirai and S. Tenpaku (2004). “Using 5 ms Segments in Concatenative Speech Synthesis”. In: *Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW)*. Pittsburgh, PA, USA, pp. 37–42 (cit. on p. 29).

G. Hofer and K. Richmond (2010). “Comparison of HMM and TMDN Methods for Lip Synchronisation”. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, pp. 454–457 (cit. on pp. 44, 47, 114).

G. Hofer, J. Yamagishi, and H. Shimodaira (2008). “Speech-driven Lip Motion Generation with a Trajectory HMM”. In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, pp. 2314–2317 (cit. on pp. 44, 47, 80, 114).

J. Hollenstein, M. Pucher, and D. Schabus (2013). “Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis”. In: *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing (AVSP)*. Annecy, France, pp. 31–36 (cit. on pp. 22, 103).

H.-W. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe (1998). “Automatic generation of synthesis units for trainable text-to-speech systems”. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. Seattle, WA, USA, pp. 293–296 (cit. on p. 29).

B. K. Horn and B. G. Schunck (1981). "Determining optical flow". In: *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203 (cit. on p. 32).

M. Hornung, F. Roitinger, and G. Zeilinger (2000). *Die österreichischen Mundarten. Eine Einführung. (The Austrian dialects. An Introduction. In German)*. Öbv&Hpt (cit. on p. 100).

F. J. Huang, E. Cosatto, and H. Graf (2002). "Triphone based unit selection for concatenative visual speech synthesis". In: *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 2. Orlando, FL, USA, pp. 2037–2040 (cit. on p. 32).

X. Huang, A. Acero, J. Adcock, H.-W. Hon, J. Goldsmith, J. Liu, and M. Plumpe (1996). "Whistler: A Trainable Text-to-Speech System". In: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*. Philadelphia, PA, USA, pp. 2387–2390 (cit. on p. 29).

A. J. Hunt and A. W. Black (1996). "Unit selection in a concatenative speech synthesis system using a large speech database". In: *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. Atlanta, GA, USA, pp. 373–376 (cit. on pp. 25, 26).

S. Imai (1983). "Cepstral analysis synthesis on the mel frequency scale". In: *Proceedings of the 1983 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 8. Boston, MA, USA, pp. 93–96 (cit. on p. 51).

Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (2001). "Investigation of State Duration Model based on Gamma distribution for HMM-based Speech Synthesis". In: *IEICE Technical Report*, vol. 101, no. 352, pp. 57–62 (cit. on p. 57).

F. Jelinek, L. R. Bahl, and R. L. Mercer (1975). "Design of a linguistic statistical decoder for the recognition of continuous speech". In: *IEEE Transactions on Information Theory*, vol. 21, no. 3, pp. 250–256 (cit. on p. 27).

J. Jiang, A. Alwan, L. E. Bernstein, P. Keating, and E. Auer (2000). "On the Correlation between Facial Movements, Tongue Movements and Speech Acoustics". In: *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. Beijing, China, pp. 42–45 (cit. on p. 110).

B.-H. Juang and L. R. Rabiner (1985). "Mixture autoregressive hidden Markov models for speech signals". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 6, pp. 1404–1413 (cit. on p. 27).

B.-H. Juang and L. R. Rabiner (1990). "The segmental K-means algorithm for estimating parameters of hidden Markov models". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 9, pp. 1639–1641 (cit. on p. 55).

V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo (2008). "The Blizzard Challenge 2008". In: *Proceedings of the Blizzard Challenge Workshop*. Brisbane, Australia, pp. 1–18 (cit. on p. 29).

H. Kawahara, J. Estill, and O. Fujimura (2001). "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT". In: *Proceedings of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*. Florence, Italy, pp. 59–64 (cit. on p. 52).

H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné (1999). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". In: *Speech Communication*, vol. 27, no. 3–4, pp. 187–207 (cit. on pp. 52, 116).

H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda (2004). "XIMERA: A New TTS from ATR Based on Corpus-Based Technologies". In: *Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW)*. Pittsburgh, PA, USA, pp. 179–184 (cit. on p. 29).

S. King and V. Karaiskos (2012). "The Blizzard Challenge 2012". In: *Proceedings of the Blizzard Challenge Workshop*. Portland, OR, USA, pp. 1–11 (cit. on p. 29).

C. J. Leggetter and P. C. Woodland (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". In: *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185 (cit. on p. 71).

B. Le Goff and C. Benoît (1996). "A Text-to-audiovisual-speech Synthesizer for French". In: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*. Philadelphia, PA, USA, pp. 2163–2166 (cit. on p. 40).

H. Leung and V. W. Zue (1984). "A procedure for automatic alignment of phonetic transcriptions with continuous speech". In: *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*. Vol. 9. San Diego, CA, USA, pp. 73–76 (cit. on p. 53).

Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang (2008). "Articulatory Control of HMM-based Parametric Speech Synthesis Driven by

Phonetic Knowledge". In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, pp. 573–576 (cit. on p. 103).

Z.-H. Ling, L. Qin, H. Lu, Y. Gao, R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen, and G.-P. Hu (2007). "The USTC and iFlytek Speech Synthesis Systems for Blizzard Challenge 2007". In: *Proceedings of the Blizzard Challenge Workshop*. Bonn, Germany, pp. 1–6 (cit. on p. 29).

Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang (2009). "Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis". In: *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185 (cit. on p. 103).

Z.-H. Ling and R.-H. Wang (2007). "HMM-Based Hierarchical Unit Selection Combining Kullback-Leibler Divergence with Likelihood Criterion". In: *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 4. Honolulu, HI, USA, pp. 1245–1248 (cit. on p. 29).

D. W. Massaro (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA, USA: MIT Press (cit. on p. 37).

D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodgriguez (1999). "Picture My Voice: Audio to Visual Speech Synthesis Using Artificial Neural Networks". In: *Proceedings of the 3rd Conference on Auditory-Visual Speech Processing (AVSP)*. Santa Cruz, CA, USA, pp. 1–6 (cit. on pp. 38, 39).

D. W. Massaro, M. M. Cohen, M. Tabain, J. Beskow, and R. Clark (2012). "Animated Speech: Research Progress and Applications". In: *Audiovisual Speech Processing*. Ed. by G. Bailly, P. Perrier, and E. Vatikiotis-Bateson. Cambridge University Press, pp. 309–345 (cit. on p. 118).

T. Masuko (2002). "HMM-Based Speech Synthesis and Its Applications". PhD thesis. Tokyo, Japan: Tokyo Institute of Technology (cit. on pp. 49, 51, 52, 62).

T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda (1998). "Text-to-visual speech synthesis based on parameter generation from HMM". In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 6. Seattle, WA, USA, pp. 3745–3748 (cit. on pp. 41, 43, 114, 115).

T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai (1996). "Speech synthesis using HMMs with dynamic features". In: *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. Atlanta, GA, USA, pp. 389–392 (cit. on p. 27).

J. J. Odell (1995). "The Use of Context in Large Vocabulary Speech Recognition". PhD thesis. Cambridge, UK: University of Cambridge (cit. on pp. 63, 64).

S. E. G. Öhman (1967). "Numerical Model of Coarticulation". In: *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320 (cit. on p. 41).

T. Okubo, R. Mochizuki, and T. Kobayashi (2006). "Hybrid Voice Conversion of Unit Selection and Generation Using Prosody Dependent HMM". In: *IEICE TRANSACTIONS on Information and Systems*, vol. E89-D, no. 11, pp. 2775–2782 (cit. on p. 29).

A. Oppenheim and R. Schafer (1968). "Homomorphic analysis of speech". In: *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226 (cit. on p. 50).

S. Ouni, M. M. Cohen, H. Ishak, and D. W. Massaro (2007). "Visual contribution to speech perception: measuring the intelligibility of animated talking heads". In: *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 47891, pp. 1–12 (cit. on p. 46).

I. S. Pandzic and R. Forchheimer, eds. (2003). *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. New York, NY, USA: John Wiley & Sons, Inc. (cit. on p. 44).

F. I. Parke (1972a). "Computer Generated Animation of Faces". MA thesis. Salt Lake Ctity, UT, USA: University of Utah (cit. on p. 36).

F. I. Parke and K. Waters (1996). *Computer facial animation*. Wellesley, MA, USA: A K Peters (cit. on p. 31).

F. I. Parke (1972b). "Computer Generated Animation of Faces". In: *Proceedings of the ACM Annual Conference*. Boston, MA, USA, pp. 451–457 (cit. on p. 36).

F. I. Parke (1974). "A Parametric Model of Human Faces". PhD thesis. Salt Lake City, UT, USA: University of Utah (cit. on pp. 36–38).

F. I. Parke (1982). "Parameterized Models for Facial Animation". In: *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 61–68 (cit. on pp. 36, 37).

A. Pearce, B. Wyvill, G. Wyvill, and D. Hill (1986). "Speech and expression: A computer solution to face animation". In: *Proceedings of Graphics Interface '86/Vision Interface '86*, pp. 136–140 (cit. on p. 37).

K. Pearson (1901). "On lines and planes of closest fit to systems of points in space". In: *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572 (cit. on pp. 32, 84).

B. Pfister and T. Kaufmann (2008). *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer Verlag (cit. on pp. 24, 25).

F. Pighin and J. P. Lewis (2006). "Facial Motion Retargeting". In: *ACM SIGGRAPH 2006 Courses, 33rd International Conference and Exhibition on Computer Graphics and Interactive Techniques*. Boston, MA, USA, pp. 1–9 (cit. on pp. 79, 80).

A. B. Poritz (1982). "Linear predictive hidden Markov models and the speech signal". In: *Proceedings of the 1982 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 7. Paris, France, pp. 1291–1294 (cit. on p. 27).

M. Pucher, F. Neubarth, E. Rank, G. Niklfeld, and Q. Guan (2003). "Combining Non-Uniform Unit Selection with Diphone Based Synthesis". In: *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*. Geneva, Switzerland, pp. 865–868 (cit. on p. 19).

M. Pucher, D. Schabus, G. Hofer, N. Kerschhofer-Puhalo, and S. Moosmüller (2012). "Regionalizing Virtual Avatars – Towards Adaptive Audio-Visual Dialect Speech Synthesis". In: *Proceedings of the 5th International Conference on Cognitive Systems (CogSys)*. Vienna, Austria, pp. 1–1 (cit. on p. 22).

M. Pucher, D. Schabus, and J. Yamagishi (2010a). "Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, pp. 2186–2189 (cit. on pp. 74, 103).

M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom (2010b). "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis". In: *Speech Communication*, vol. 52, no. 2, pp. 164–179 (cit. on pp. 74, 116).

L. R. Rabiner (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286 (cit. on pp. 49, 56).

L. Revéret, G. Bailly, and P. Badin (2000). "MOTHER: A New Generation of Talking Heads Providing a Flexible Articulatory Control for Video-Realistic Speech Animation". In: *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. Beijing, China, pp. 755–758 (cit. on p. 41).

K. Richmond, P. Hoole, and S. King (2011). "Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus".

In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. http://www.mngu0.org. Florence, Italy, pp. 1505–1508 (cit. on p. 99).

J. Rissanen (1978). "Modeling by shortest data description". In: *Automatica*, vol. 14, no. 5, pp. 465–471 (cit. on p. 66).

S. Rouibia and O. Rosec (2005). "Unit Selection for Speech Synthesis Based on a New Acoustic Target Cost". In: *Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH/INTERSPEECH)*. Lisbon, Portugal, pp. 2565–2568 (cit. on p. 29).

Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura (1992). "ATR $\nu$-Talk Speech Synthesis System". In: *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP)*. Banff, AB, Canada, pp. 483–486 (cit. on p. 25).

K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda (2006). "An HMM-based singing voice synthesis system". In: *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH)*. Pittsburgh, PA, USA, pp. 2274–2277 (cit. on p. 114).

S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (2000). "HMM-based text-to-audio-visual speech synthesis". In: *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. Vol. 3. Beijing, China, pp. 25–28 (cit. on pp. 33, 34, 114, 115, 122).

D. Schabus (2009). "Interpolation of Austrian German and Viennese Dialect/Sociolect in HMM-based Speech Synthesis". MA thesis. Vienna, Austria: Vienna University of Technology (cit. on pp. 19, 23).

D. Schabus, M. Pucher, and G. Hofer (2011). "Simultaneous speech and animation synthesis". In: *ACM SIGGRAPH Posters, 38th International Conference and Exhibition on Computer Graphics and Interactive Techniques*. Vancouver, BC, Canada, 8:1–8:1 (cit. on pp. 21, 122, 144).

D. Schabus, M. Pucher, and G. Hofer (2012a). "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 3313–3316 (cit. on pp. 21, 75, 84, 99, 100, 144, 145).

D. Schabus, M. Pucher, and G. Hofer (2012b). "Speaker-adaptive visual speech synthesis in the HMM-framework". In: *Proceedings of the 13th Annual Conference of the International Speech Communication Associa-*

*tion (INTERSPEECH)*. Portland, OR, USA, pp. 979–982 (cit. on pp. 21, 135, 144).

D. Schabus, M. Pucher, and G. Hofer (2013). "Objective and Subjective Feature Evaluation for Speaker-Adaptive Visual Speech Synthesis". In: *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing (AVSP)*. Annecy, France, pp. 37–42 (cit. on pp. 21, 75, 84, 87, 144).

D. Schabus, M. Pucher, and G. Hofer (2014a). "Joint Audiovisual Hidden Semi-Markov Model-based Speech Synthesis". In: *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347 (cit. on pp. 21, 47, 99, 102, 113, 143–145).

D. Schabus, M. Pucher, and P. Hoole (2014b). "The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland, pp. 3411–3416 (cit. on pp. 21, 75, 99, 145).

K. Shinoda and T. Watanabe (2000). "MDL-based context-dependent subword modeling for speech recognition". In: *Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 2, pp. 79–86 (cit. on p. 66).

J. Shlens (2014). *A Tutorial on Principal Component Analysis*. URL: `http://arxiv.org/abs/1404.1100` (cit. on pp. 32, 84, 86).

S. S. Stevens, J. Volkmann, and E. B. Newman (1937). "A Scale for the Measurement of the Psychological Magnitude Pitch". In: *The Journal of the Acoustical Society of America*, vol. 8, pp. 185–190 (cit. on p. 50).

W. H. Sumby and I. Pollack (1954). "Visual Contribution to Speech Intelligibility in Noise". In: *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215 (cit. on pp. 30, 46).

M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi (2006). "A Style Adaptation Technique for Speech Synthesis Using HSMM and Suprasegmental Features". In: *IEICE Transactions on Information and Systems*, vol. E89-D, no. 3, pp. 1092–1099 (cit. on p. 74).

M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi (1999). "Text-to-Audio-Visual Speech Synthesis Based on Parameter Generation from HMM". In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*. Budapest, Hungary, pp. 959–962 (cit. on pp. 42, 47, 114).

M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda (1998a). "Visual Speech Synthesis Based On Parameter Generation From HMM: Speech-Driven And Text-And-Speech-Driven Approaches". In: *Proceedings of*

*the 2nd International Conference on Auditory-Visual Speech Processing (AVSP)*. Terrigal, Sydney, Austraila, pp. 221–226 (cit. on pp. 42–44, 80, 114, 115).

M. Tamura, T. Masuko, K. Tokuda, and T. Kobayash (1998b). "Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR". In: *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW)*. Jenolan, Australia, pp. 273–276 (cit. on p. 72).

J. Tao, L. Xin, and P. Yin (2009). "Realistic Visual Speech Synthesis Based on Hybrid Concatenation Method". In: *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 469–477 (cit. on pp. 44, 80).

L. Terry (2011). "Audio-visual asynchrony modeling and analysis for speech alignment and recognition". PhD thesis. Evanston, IL, USA: Northwestern University (cit. on p. 114).

B.-J. Theobald, J. A. Bangham, I. A. Matthews, and G. C. Cawley (2002). "Towards video realistic synthetic visual speech". In: *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 4. Orlando, FL, USA, pp. 3892–3895 (cit. on p. 33).

B.-J. Theobald, J. A. Bangham, I. A. Matthews, and G. C. Cawley (2004). "Near-videorealistic synthetic talking faces: implementation and evaluation". In: *Speech Communication*, vol. 44, no. 1–4. Special Issue on Audio Visual speech processing, pp. 127–140 (cit. on pp. 33, 34).

B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei (2008). "LIPS2008: visual speech synthesis challenge". In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, pp. 2310–2313 (cit. on p. 129).

K. Tokuda, T. Kobayashi, and S. Imai (1995). "Speech parameter generation from HMM using dynamic features". In: *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. Detroit, MI, USA, pp. 660–663 (cit. on pp. 27, 28, 58–60, 70, 116).

K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai (1994). "Mel-Generalized Cepstral Analysis - A Unified Approach to Speech Spectral Estimation". In: *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)*. Yokohama, Japan, pp. 1043–1046 (cit. on pp. 50, 51, 67).

K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi (1999). "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling". In: *Proceedings of the 1999 IEEE International Con-*

*ference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. Phoenix, AZ, USA, pp. 229–232 (cit. on pp. 27, 60, 62).

K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura (2013). "Speech Synthesis Based on Hidden Markov Models". In: *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252 (cit. on pp. 26, 28, 49, 55, 60, 65).

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura (2000). "Speech parameter generation algorithms for HMM-based speech synthesis". In: *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 3. Istanbul, Turkey, pp. 1315–1318 (cit. on p. 70).

M. Toman, M. Pucher, and D. Schabus (2013a). "Cross-variety speaker transformation in HSMM-based speech synthesis". In: *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*. Barcelona, Spain, pp. 77–81 (cit. on p. 74).

M. Toman, M. Pucher, and D. Schabus (2013b). "Multi-variety adaptive acoustic modeling in HSMM-based speech synthesis". In: *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*. Barcelona, Spain, pp. 83–87 (cit. on pp. 74, 100).

C. Valentini-Botinhao, M. Toman, M. Pucher, D. Schabus, and J. Yamagishi (2014). "Intelligibility Analysis of Fast Synthesized Speech". In: *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore, pp. 2922–2926 (cit. on pp. 57, 103).

A. J. Viterbi (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269 (cit. on p. 25).

L. Wang, W. Han, F. K. Soong, and Q. Huo (2011a). "Text Driven 3D Photo-Realistic Talking Head". In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Florence, Italy, pp. 3307–3308 (cit. on pp. 34, 36).

L. Wang, X. Qian, W. Han, and F. K. Soong (2010). "Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, pp. 446–449 (cit. on pp. 34, 35, 47, 80).

L. Wang, J. Wu, X. Zhuang, and F. Soong (2011b). "Synthesizing visual speech trajectory with minimum generation error". In: *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal*

*Processing (ICASSP)*. Prague, Czech Republic, pp. 4580–4583 (cit. on pp. 34, 45, 47, 114).

P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young (1994). "Large vocabulary continuous speech recognition using HTK". In: *Proceedings of the 1994 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2. Adelaide, Australia, pp. 125–128 (cit. on p. 27).

A. Wrench (1999). *The MOCHA-TIMIT articulatory database*. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (cit. on p. 99).

J. Yamagishi (2006). "Average-Voice-Based Speech Synthesis". PhD thesis. Tokyo, Japan: Tokyo Institute of Technology (cit. on p. 49).

J. Yamagishi and T. Kobayashi (2007). "Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training". In: *IEICE Transactions on Information and Systems*, vol. E90-D, no. 2, pp. 533–543 (cit. on pp. 72, 74, 136).

J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai (2009a). "Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm". In: *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83 (cit. on pp. 74, 136).

J. Yamagishi, T. Masuko, and T. Kobayashi (2004). "A study on state duration modeling using lognormal distribution for HMM-based speech synthesis". In: *Proceedings of the 2004 Spring Meeting of the Acoustical Society of Japan (ASJ)*. 1-7-7, pp. 225–226 (cit. on p. 57).

J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals (2009b). "Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis". In: *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230 (cit. on pp. 73, 74).

J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi (2003). "Modeling of various speaking styles and emotions for HMM-based speech synthesis". In: *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*. Geneva, Switzerland, pp. 2461–2464 (cit. on p. 74).

J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi (2002). "A Context Clustering Technique for Average Voice Model in HMM-Based Speech Synthesis". In: *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH)*. Denver, CO, USA, pp. 133–136 (cit. on p. 74).

J. Yamagishi and O. Watts (2010). "The CSTR/EMIME HTS System for Blizzard Challenge 2010". In: *Proceedings of the Blizzard Challenge Workshop*. Kansai Science City, Japan, pp. 1–6 (cit. on pp. 115, 116).

J.-H. Yang, Z.-W. Zhao, Y. Jiang, G.-P. Hu, and X.-R. Wu (2006). "Multi-tier Non-uniform Unit Selection for Corpus-based Speech Synthesis". In: *Proceedings of the Blizzard Challenge Workshop*. Pittsburgh, PA, USA, pp. 1–4 (cit. on p. 29).

H. Yehia, P. Rubin, and E. Vatikiotis-Bateson (1998). "Quantitative association of vocal-tract and facial behavior". In: *Speech Communication*, vol. 26, no. 1–2, pp. 23–43 (cit. on p. 110).

T. Yoshimura (2002). "Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-Based Text-to-Speech Systems". PhD thesis. Nagoya, Japan: Nagoya Institute of Technology (cit. on pp. 49, 66).

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (1998). "Duration modeling for HMM-based speech synthesis". In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, pp. 29–32 (cit. on pp. 27, 56, 120).

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (1999). "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis". In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*. Budapest, Hungary, pp. 2374–2350 (cit. on pp. 27, 58, 59).

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (2000). "Speaker interpolation for HMM-based speech synthesis system". In: *Journal of the Acoustical Science of Japan (E)*, vol. 21, no. 4, pp. 199–206 (cit. on p. 28).

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (2001). "Mixed Excitation for HMM-based Speech Synthesis". In: *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH/INTERSPEECH)*. Aalborg, Denmark, pp. 2263–2266 (cit. on p. 71).

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland (2006). *The HTK Book (for HTK version 3.4)*. Cambridge, UK: Cambridge University Engineering Department (cit. on pp. 27, 105).

S.-Z. Yu (2010). "Hidden semi-Markov models". In: *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243 (cit. on pp. 56, 57).

H. Zen (2006). "Reformulating HMM as a Trajectory Model by Imposing Explicit Relationships between Static and Dynamic Features". PhD thesis. Nagoya, Japan: Nagoya Institute of Technology (cit. on pp. 49, 70).

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda (2007a). "The HMM-based Speech Synthesis System (HTS) Version 2.0". In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW)*. Bonn, Germany, pp. 294–299 (cit. on pp. 27, 60, 65).

H. Zen, T. Toda, M. Nakamura, and K. Tokuda (2007b). "Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005". In: *IEICE Transactions on Information and Systems*, vol. E90-D, no. 1, pp. 325–333 (cit. on p. 52).

H. Zen, K. Tokuda, and A. W. Black (2009). "Statistical parametric speech synthesis". In: *Speech Communication*, vol. 51, no. 11, pp. 1039–1064 (cit. on pp. 26, 28, 49, 53).

H. Zen, K. Tokuda, and T. Kitamura (2007c). "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences". In: *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173 (cit. on pp. 60, 70).

H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (2004). "Hidden Semi-Markov Model Based Speech Synthesis". In: *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*. Jeju, South Korea, pp. 1185–1188 (cit. on pp. 57, 116).

H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura (2007d). "A hidden semi-Markov model-based speech synthesis system". In: *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 825–834 (cit. on pp. 57, 115).