Visio-articulatory to acoustic conversion of speech

Michael Pucher, Dietmar Schabus
The Telecommunications Research Center Vienna (FTW)*

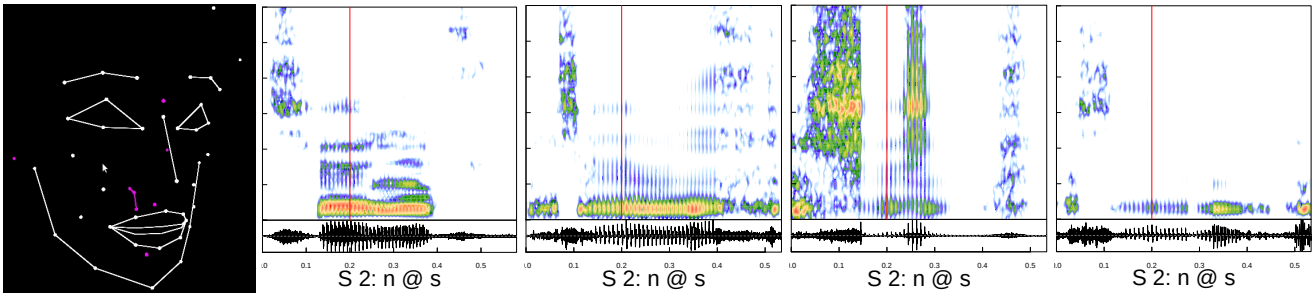

Figure 1: *Visio-articulatory recordings (1), original audio recording (2), visio-articulatory audio reconstruction (3), articulatory only audio reconstruction (4), visual only audio reconstruction (5) for the German word "schönes" [ʃ ʌː n ə s].*

1 Introduction

In this paper we evaluate the performance of combined visual and articulatory features for the conversion to acoustic speech. Such a conversion has possible applications in silent speech interfaces, which are based on the processing of non-acoustic speech signals. With an intelligibility test we show that the usage of joint visual and articulatory features can improve the reconstruction of acoustic speech compared to using only articulatory or visual data. An improvement can be achieved when using the original or using no voicing information.

2 Visio-articulatory-acoustic data

We have recorded a 30-year old male native speaker of Austrian German reading 320 phonetically diverse sentences off a computer screen. Facial movement was recorded using a NaturalPoint OptiTrack Expression system using seven FLEX:V100R2 infrared cameras. This system records the 3D position of 37 reflective markers glued to the speakers face at 100 Hz. Articulatory movement was recorded with a Carstens Medizinelektronik Articulograph AG501 (Carstens Medizinelektronik, 2014) EMA system. A detailed description and analysis of the recordings is given in [Schabus et al. 2014].

3 Conversion method

For voice conversion we first train a Gaussian mixture model (GMM) for the joint vector (x_t, y_t), where x are the Principal Component Analysis (PCA) reduced visual (VIS), articulatory (EMA), or visio-articulatory (VIS_EMA) features with dimension 30, 10, and 40 and y are the Mel-Cepstral acoustic features (MFCC).

As described in [Toda et al. 2007] the conversion based on the minimum Mean Squared Error (MSE) criterion is defined as

$$\hat{y_t} = E[y_t|x_t] = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) E_{m,t}^{(y)} \quad (1)$$

where $\hat{y_t}$ are the predicted acoustic features, M are the GMM mixture components (=128), $\lambda^{(z)}$ are the GMM parameters and $E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} (x_t - \mu_m^x)$ with $\Sigma_m^{(yx)}$ and $\Sigma_m^{(xx)^{-1}}$ being full co-variance matrices.

4 Experiments

For the experiments we used the training data to train 3 GMMs, one for articulatory, visual, and spectral features (VIS_EMA_MFCC), one for articulatory and spectral features (EMA_MFCC), and one for visual and spectral features (VIS_MFCC). The method described in Section 3 is then used to reconstruct the acoustic spectral features using the GMM for 28 test sentences not contained in the training data. The fundamental frequency (F0) is taken from the original recordings or no F0 is provided. This results in 6 different models that are evaluated. The wav files are then synthesized from the reconstructed spectral features and the F0 files.

6 listeners had to listen to the samples synthesized with the different methods. In an intelligibility test they had to write down the words that they understood. Each sentence was presented to each listener only with one method, but listeners were allowed to listen to the sentence as often as they liked.

Table 1: *Word-error-rate (WER) for the different methods in %.*

	Original F0	No F0
VIS_EMA_MFCC	75.5	84.4
EMA_MFCC	93.2	99.3
VIS_MFCC	95.2	98.0

Table 1 shows the results of the intelligibility test where we can see that the usage of joint visio-articulatory features (VIS_EMA_MFCC) can improve the recognition performance compared to using articulatory or visual features only. In future work we want to add additional visual features derived from the recorded video data.

References

SCHABUS, D., PUCHER, M., AND HOOLE, P. 2014. The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech. In *LREC 2014*, 3411–3416.

TODA, T., BLACK, A., AND TOKUDA, K. 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on 15*, 8 (Nov), 2222–2235.

*e-mail:{pucher, schabus}@ftw.at