Exploiting visual information for NAM recognition

Panikos Heracleous, Denis Beautemps, Viet-Anh Tran, Helene Loevenbruck, and Gérard Bailly
GIPSA-lab, Speech and Cognition Department
CNRS UMR 5216 / Stendhal University / UJF / INPG
961 rue de la Houille Blanche Domaine universitaire BP 46
F - 38402 Saint Martin d'Hères cedex , France

{*Panikos.Heracleous, Denis.Beautemps, Viet-Anh.Tran, Helene.Loevenbruck, Gerard.Bailly*}*@gipsa-lab.inpg.fr*

Abstract: Non-audible murmur (NAM) is an unvoiced speech received through body tissue using special acoustic sensors (i.e., NAM microphones) attached behind the talkers ear. Although NAM has different frequency characteristics compared to normal speech, it is possible to perform automatic speech recognition (ASR) using conventional methods. In using a NAM microphone, body transmission and the loss of lip radiation act as a low-pass filter; as a result, higher frequency components are attenuated in NAM signal. A decrease in NAM recognition performance is attributed to spectral reduction. To address the problem of loss of lip radiation, visual information extracted from the talker's facial movements is fused with NAM speech. Experimental results revealed a relative improvement of 39% when fused NAM speech and facial information were used as compared to using only NAM speech. Results also showed that improvements in the recognition rate depend on the place of articulation.
Keywords: NAM, speech recognition, facial movements, fusion.
Classification: Science and engineering for electronics

References

[1] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, *Non-Audible Murmur Recognition*, in Proceedings of EUROSPEECH, pp. 2601–2604, 2003.

[2] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and Z. Huang, *Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement*, in Proceedings of ASRU, pp. 249–253, 2003.

[3] S. C. Jou, T. Schultz, and A. Weibel, *Adaptation for Soft Whisper Recognition Using a Throat Microphone*, in Proceedings of Interspeech2004-ICSLP, pp. 1493–1496, 2004.

[4] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano, *Non-Audible Murmur (NAM) Recognition Using a Stethoscopic NAM microphone*, In Proceedings of Interspeech2004-ICSLP, pp. 1469–1472, 2004.

[5] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, *Investigating the Role of the Lombard Reflex in Non-Audible Murmur (NAM) Recognition*, in Proceedings of Interspeech2005-EUROSPEECH, pp. 2649–2652, 2005.
[6] L. Revéret, and C. Benoît, *A new 3D lip model for analysis and synthesis of lip motion in speech production*, in Proceedings of AVSP, 1998.
[7] L. Revéret, G. Bailly, and P. Badin, *MOTHER: a new generation of talking heads providing a flexibile articulatory control for video-realistic speech animation*, in Proceeding os ICSLP, China, 755-758, 2000.
[8] H. Bourlard ,and S. Dupont, *A new ASR approach based on independent processing and recombination of partial frequency bands*, in Proceedings of International Conference on Spoken Language Processing, pp 426-429, 1996.
[9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, *Recent advances in the automatic recognition of audiovisual speech*, in Proceedings of the IEEE, vol. 91, Issue 9, pp. 1306–1326, 2003.

1 Introduction

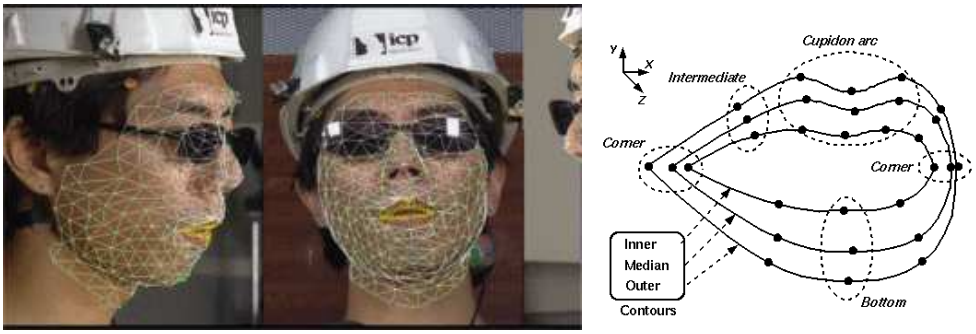
NON-audible murmur (NAM) is a very quietly uttered speech received through body tissue. A special acoustic sensor (i.e., NAM microphone) is attached behind the talkers ear and receives very quiet sounds that are inaudible to other listeners near the talker [1]. Such approaches have been introduced in [2, 3] for speech enhancement or speech recognition.

Similar to whisper speech, NAM is unvoiced speech produced by vocal cords that do not vibrate and incorporate any fundamental (F0) frequency. Moreover, body tissue and loss of lip radiation act as a low-pass filter that attenuates high-frequency components. However, experimental results showed that the NAM spectral components still provide sufficient information to distinguish and recognize sounds accurately. The authors have reported experiments for NAM recognition with very promising results. With a small amount of adaptation data, 93.9% word accuracy for a 20k Japanese vocabulary dictation task was achieved [4]. Moreover, the authors conducted experiments using simulated and real noisy test data with clean training data to prove the noise robustness of NAM microphones. Although NAM microphones are highly robust against noise when using simulated noisy data, their performance decreased with real noise data because of the effect of Lombard reflex [5].

This article focuses on audio-visual NAM recognition. To address the loss of lip radiation, visual information extracted from the talker's facial movements were integrated with NAM speech using concatenative feature fusion and multi-stream HMM decision fusion and the joint feature vectors were used for NAM recognition. When visual information was integrated, a significant improvement in recognition accuracy was obtained as compared to using acoustic parameters only. Another issue was the relationship between accuracy and place of the articulation of the Japanese consonants. Recognition accuracy for consonants articulated at the front (e.g., bilabial consonants)

have shown significantly higher improvement compared with consonants articulated at the back (e.g., velars).

2 Facial shape parameters extraction



(a) (b)

Fig. 1: Characteristic points used for capturing facial movements.

The face and profile views of the subject have been filmed under good lighting conditions. The system captured at a sampling rate of 50 Hz the three-dimensional (3-D) positions of 112 colored beads glued on the speakers face (Fig. **1**a) in synchrony with the acoustic signal sampled at 16000 Hz. Fig. **1**b shows the collection of 30 lip points using a generic 3D geometric model of the lips [6].

The shape model is built using Principal Component Analysis (PCA). Successive applications of PCA performed on selected subsets of the data generate the main directions retained as linear predictors for the whole data set [7]. The mobile points P of the face deviate from their average position B by a linear composition of basic components M loaded by factors α (articulatory parameters).

$$P = B + \alpha M \tag{1}$$

We only used first 5 parameters of extracted 12 linear components M that explain more than 90% of the data variance using the following iterative linear predictions on data residual: the first component of the PCA on the lower teeth (LT) values leads to the first jaw predictor. PCA on the residual lips values (without jaw1's influence) usually gives 3 pertinent lip predictors (e.g., lip protrusion; lip closing, mainly required for bilabials; lip raising, mainly required for labiodental fricatives). Movements of the throat linked the underlying movements of the larynx and the hyoid bone, serves as fifth one. The resulting articulatory model also includes others components for head movements and facial expressions but here we consider only these 5 components related to speech articulation. Then, to synchronize with the audio, video parameters were interpolated at 200 Hz.

3 NAM and lip shape fusion

Audio-visual fusion [9] is the integration of available single modality streams into a combined one. In ASR, visual information is also used to improve the recognition performance in noisy environments. In this article, facial information seeks to address the problem of loss of lip radiation during the production of NAM speech. In this work, audio and video streams were used. Our aim was to combine the two streams into a bimodal one and use the joint audio-visual feature vectors in the HMM system to gain higher performance when compared to single modality streams.

3.1 Concatenative feature fusion

Concatenative feature fusion is maybe the simplest audio-visual fusion approach. The feature concatenation uses the concatenation of the synchronous audio and visual features as the joint bimodal feature vector

$$O_t^{AV} = [O_t^{(A)^T}, O_t^{(V)^T}]^T \in R^D \tag{2}$$

where O_t^{AV} is the joint audio-visual feature vector, $O_t^{(A)}$ the audio feature vector, $O_t^{(V)}$ the visual feature vector, and D the dimensionality of the joint feature vector.

3.2 Multi-stream HMM decision fusion

Decision fusion captures the reliability of each stream by combining the likelihoods of single-modality HMM classifiers. Such an approach has been used in multi-band audio-only ASR [8] and in audio-visual speech recognition [9]. The emission likelihood of multi-stream HMM is the product of emission likelihoods of single-modality components weighted appropriately by stream weights. Given the O bimodal observation vector, i.e., NAM speech and facial modality, the emission probability of multi-stream HMM is given by

$$b_j(O_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm})]^{\lambda_{sjt}} \tag{3}$$

where $N(O; \mu, \Sigma)$ is the value in O of a multivariate Gaussian with mean μ and covariance matrix Σ. For each stream s, M_s Gaussians in a mixture are used, each weighted with c_{jsm}. The contribution of each stream is weighted by λ_{sjt}. In this study, we assume that the stream weights do not depend on state j and time t. It is assumed, however that

$$0 \leq \lambda_s, \lambda_f \leq 1 \tag{4}$$

and

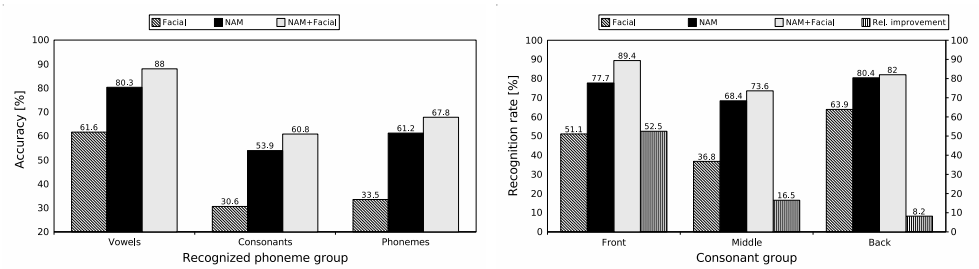
$$\lambda_s + \lambda_s = 1 \tag{5}$$

where λ_s is the NAM speech stream weight, and λ_f is the facial stream weight. Since those weights cannot be obtained by Maximum Likelihood Estimation (MLE), they were adjusted manually to 0.7 and 0.3 values, respectively, by maximizing the accuracy on a held-out data.

4 Experiments and results

In this section, recognition experiments are introduced for Japanese NAM recognition by using visual information. In the experiments, 5 vowel and 17 consonant context-independent HMMs were trained. Each HMM state was modeled with 16 Gaussian mixtures. The data used were 212 Japanese continuous utterances, which contained 2568 vowels and 2409 consonants, respectively. A total of 3446 phonemes (1785 vowels and 1661 consonants) were used for training, while 1531 phonemes (783 vowels and 748 consonants) were used for testing. The continuous utterances were hand labeled at phonemic level, and then the phonemes were extracted from the utterances. The acoustic parameter vectors were of length 36 (12 MFCC, 12 $\Delta MFCC$ and 12 $\Delta\Delta MFCC$). The dimension of the visual stream was 15 (5 PCA visual components, first and second derivatives). Therefore, the dimension of the joint audio-visual feature vectors was 51. This size was acceptable and no further processing was applied to reduce the dimensionality. The HTK3.1 Toolkit was used for training and testing.

4.1 NAM phoneme recognition



(a) (b)

Fig. 2: Phoneme recognition using concatenative feature fusion of NAM and facial information.

Fig. **2**a shows the obtained results for vowel, consonant and phoneme recognition when concatenative fusion was applied. In the case of vowel recognition, the accuracy obtained by using the visual information was 88%, showing a relative improvement of 39% when compared with using acoustic information only. In the case of consonant recognition, the relative improvement was 15%, whereas in all-phoneme recognition the relative improvement was 17%. A comparatively lower improvement in consonants was obtained as vowels contain more facial information. However, some consonants (e.g. velars) are articulated at the back and the contained facial information is low. The obtained results showed the effectiveness of integrating visual information in NAM recognition. On average, 23.6% relative improvement was achieved, which is a promising result.

4.2 Consonant recognition considering the place of articulation

To investigate the role of articulation in audio-visual NAM recognition, an experiment was conducted for the recognition of consonants, taking into account the place of the articulation. Based on the International Phonetic Alphabet IPA, the consonants were classified into three groups considering the place of articulation, as follows:

- Front consonants: Bilabials.
- Middle consonants: Alveolar, and post alveolar.
- Back consonants: Palatal, velar and glottal.

Three HMM sets were created for each consonant group respectively using the appropriate data. Fig. **2**b shows the obtained results. The improvement in accuracy depended on the place of the articulation: the relative improvement in the case of the front consonants was 52.5%, whereas for middle and back consonants it decreased to 16.5% and 8.2%, respectively.

4.3 Comparing concatenative feature fusion and multi-stream HMM decision fusion

The presented results were obtained using concatenative feature fusion. Recognition accuracies were further improved using multi-stream HMM decision fusion. In almost all the cases, higher accuracy was achieved using multi-stream HMM decision fusion. There was a relative improvement of 14.5% on average, which showed the efficacy of the method. As described earlier, decision fusion captures the reliability of each stream by combining the likelihoods of single-modality HMM classifiers.

5 Conclusion

Audio-visual NAM recognition was presented. In the case of NAM production, the loss of lip radiation and body transmission act as a low-pass filter attenuating higher frequency components. To deal with the loss of lip radiation, visual information extracted from the talker's facial movements were fused with the acoustic modality using concatenative feature fusion and multi-stream HMM fusion, and audio-visual NAM phoneme recognition experiments were carried out. The results were promising and demonstrated the effectiveness of the proposed method. The experimental results showed that the accuracy improvement achieved by using visual information also highly depends on the place of the articulation of the phonemes. As future work, we plan to extend the recognition task to Large Vocabulary Continuous Speech Recognition (LVCSR) systems.

Acknowledgments

The authors would like to thank Professor Kiyohiro Shihano and the members of the Speech and Acoustics Processing Laboratory, Nara Institute of Science and Technology, Japan for providing the NAM microphones and helping in recording Japanese NAM data.