EXPLOITING MULTIMODAL DATA FUSION IN ROBUST SPEECH RECOGNITION

Panikos Heracleous[1], *Pierre Badin*[2], *Gérard Bailly*[2], *and Norihiro Hagita*[1]

[1]ATR, Intelligent Robotics and Communication Laboratories, Japan
[2]GIPSA-lab, Speech and Cognition Department, UMR 5216, CNRS-Grenoble University, France
E-mail:panikos@atr.jp

ABSTRACT

This article introduces automatic speech recognition based on Electro-Magnetic Articulography (EMA). Movements of the tongue, lips, and jaw are tracked by an EMA device, which are used as features to create Hidden Markov Models (HMM) and recognize speech only from articulation, that is, without any audio information. Also, automatic phoneme recognition experiments are conducted to examine the contribution of the EMA parameters to robust speech recognition. Using feature fusion, multistream HMM fusion, and late fusion methods, noisy audio speech has been integrated with EMA speech and recognition experiments have been conducted. The achieved results show that the integration of the EMA parameters significantly increases an audio speech recognizer's accuracy, in noisy environments.

1. INTRODUCTION

Speech is the most natural form of communication for human beings and is often described as a uni-modal communication channel. However, it is well known that speech is multi-modal in nature and includes the auditive, visual, and tactile modalities. Other less natural modalities such as electromyographic signal, invisible articulator display, or brain electrical activity or electromagnetic activity can also be considered. Therefore, in situations where audio speech is not available or is corrupted because of disability or adverse environmental condition, people may resort to alternative methods such as augmented speech.

In several automatic speech recognition systems, visual information from lips/mouth and facial movements has been used in combination with audio signals. In such cases, visual information is used to complement the audio information to improve the system's robustness against acoustic noise [1].

For the orally educated deaf or hearing-impaired people, lip reading remains a crucial speech modality, though it is not sufficient to achieve full communication. Therefore, in 1967, Cornett developed the Cued Speech system as a supplement to lip reading [2]. Recently, studies have been presented on automatic Cued Speech recognition using hand gestures in combination with lip/mouth information [3].

Several other studies have been introduced that deal with the problem of alternative speech communication based on speech modalities other than audio speech. A method for communication based on inaudible speech received through body tissues has been introduced using the Non-Audible Murmur (NAM) microphone. NAM microphones have been used for receiving and automatically recognizing sounds of speech-impaired people, for ensuring privacy in communication, and for achieving robustness against noise [4, 5]. Aside from automatic recognition of NAM speech, silicon NAM microphones were used for NAM-to-speech conversion [6, 7]

A few researchers have addressed the problem of augmented speech based on the activation signal of the muscles produced during speech production [8]. The OUISPER project [9] attempts to automatically recognize and resynthesize speech based on the signals of tongue movements captured by an ultrasound device in combination with lip information.

The present study aims to assess the possibility of developing automatic speech recognition based on articulatory information only, that is, without any audio information. It also aims to quantify the contribution of the tongue, which is usually an invisible articulator, in comparison with that of the lips. An EMA device was used to track the movements of the tongue, jaw, and lips during speech production. These parameters were used as features to create HMMs, and automatic phoneme recognition experiments were conducted. Similar studies dealing with the automatic recognition of articulatory speech in the English lan-

(a) (b)

Fig. 1: (a) Photo of EMA receptor coils attached to the subject's tongue; (b) Locations of the six active coils (white disks with black centers) in the midsagittal plane, superposed for ease of interpretation on a MRI image where the speech articulators have been outlined.

This work was supported by KAKENHI 21118003 project.

guage have been introduced in [10, 11, 12]. This article, however, focuses on the contribution of the EMA parameters to an automatic recognition system's robustness against noise in the French language, which shows differences in articulation from the English language.

2. METHODOLOGY

2.1. Extracting the articulatory features

For tracking of articulatory movements, Electro-magnetic Articulography (EMA) [13] presents a good compromise: the Carstens AG100 used in the present study can simultaneously track the vertical and horizontal coordinates in the midsagittal plane of 10 receiver coils that can be glued to the various oro-facial articulators inside and outside the vocal tract. The sampling frequency was 500 Hz, and the accuracy of the system was better than 0.1 cm. The coils have the advantage of tracking flesh points, i.e. physical locations of the articulators, in contrast to the medical imaging techniques that provide only contours.

A drawback of this technique is the poor spatial resolution related to the limited number of points. However, it has been shown that the number of degree of freedom of articulators for speech (jaw, lips, tongue, velum) is limited, and that a small but sufficient number of locations can allow to retrieve measurements with a good accuracy [14]. Finally, another important drawback of EMA is its partially invasive nature: the receiver coils have a diameter of about 0.3 cm and must be connected to the device by thin wires that can slightly interfere with the articulation.

The authors used six coils of the AG100, as illustrated in Fig. 1b; a jaw coil was attached to the lower incisors, whereas a tip coil, a mid coil, and a back coil were respectively attached at approximately 1.2 cm, 4.2 cm, and 7.3 cm from the extremity of the tongue; an upper lip coil and a lower lip coil were attached to the boundaries between the vermilion and the skin in the midsagittal plane. Another two coils attached to the upper incisor and to the nose served as reference for alignment. The audio-speech signal was recorded at a sampling frequency of 22050 Hz, in synchronization with the EMA parameters, which were recorded at a sampling frequency of 500 Hz.

2.2. Corpus and HMM modeling

The corpus consisted of a set of two repetitions of 224 nonsense vowel-consonant-vowel (VCV) sequences (slow speech, where, C is one of the 16 French consonants and V is one of the 14 French oral and nasal vowels); two repetitions of 109 pairs of consonant-vowel-consonant (CVC) real French words, minimal pairs differing only by one phoneme (the French version of the Diagnostic Rhyme Test); 68 short French sentences; and 9 longer French sentences. The continuous sentences were used in order to increase the training data. The corpus contained 4081 allophones (i.e., 40% VCV, 30% CVC, and 30% from continuous sentences). For HMM training and test, 2721 (i.e., two-third) and 1360 (i.e., one-third) of the phones were used, respectively. The test data contained 682 vowel instances and 568 consonant instances. The phoneme instances were extracted from the sentences using a forced alignment based on the audio signal, followed by a manual correction of the segmentations, and the training and test utterances consisted of isolated phones.

For HMM modeling, 38 context-independent, left-to-right with no skip, 3-state phoneme HMMs were used. Eight Gaussians per state and a diagonal covariance matrix were used. The audio signal was down-sampled to 16000 Hz; subsequently, 12 Mel-Frequency Cepstral Coefficients (MFCC), along with the first and second derivatives, were extracted. The EMA signal was down-sampled to 100 Hz in order to be synchronized with the audio feature extraction rate (i.e., 10 ms). Because the EMA coordinates were partially correlated, a global Principal Component Analysis (PCA) was applied before HMM modeling. Three articulatory HMM sets were trained using all the PCA components, along with the first and second derivatives. In the first HMM set, the coordinates of lips and jaw (LJ) were used (i.e., 6 PCA components, first and second derivatives). In the second HMM set, the parameters of tongue (T) were used (i.e., 6 PCA components, first and second time derivatives). Finally, a common HMM set for lips, jaw, and tongue (LJT) was created (i.e., 6 lips PCA components, first and second derivatives; 6 tongue PCA components, first and second derivatives).

2.3. Fusion methods

In this section, the fusion methods used to integrate the audio signal with the EMA signal are introduced. Specifically, in this study a feature fusion method and two decision methods were used; a state synchronous and a state asynchronous fusion methods.

2.3.1. Concatenative feature fusion

The feature concatenation is the simplest state synchronous fusion method. It uses the concatenation of the synchronous audio speech signal and EMA signal as the joint feature vector

$$O_t^{AE} = [O_t^{(A)^T}, O_t^{(E)^T}]^T \in R^D \quad (1)$$

where O_t^{AE} is the joint audio-EMA feature vector, $O_t^{(A)}$ the audio feature vector, $O_t^{(E)}$ the EMA feature vector, and D the dimension of the joint feature vector. In these experiments, the dimension of the audio stream was 36 and the dimension of the EMA stream was also 36. The dimension D of the joint audio-EMA feature vectors was, therefore 72.

2.3.2. Multistream HMM fusion

Multistream HMM fusion is a state synchronous decision fusion, which captures the reliability of each stream, by combining the likelihoods of single-stream HMM classifiers [1, 15]. The emission likelihood of multistream HMM is the product of

the emission likelihoods of single-stream components weighted appropriately by stream weights. Given the O combined observation vector, that is, the audio and EMA elements, the emission probability of multistream HMM is given by

$$b_j(O_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm})]^{\lambda_s} \quad (2)$$

where $N(O; \mu, \Sigma)$ is the value in O of a multivariate Gaussian with mean μ and covariance matrix Σ, and S is the number of the streams. For each stream s, M_s Gaussians in a mixture are used, with each weighted with c_{jsm}. The contribution of each stream is weighted by λ_s. In this study, we assume that the stream weights do not depend on state j and time t. However, two constraints were applied. Namely,

$$0 \leq \{\lambda_a, \lambda_e\} \leq 1, \quad and \quad \lambda_a + \lambda_e = 1 \quad (3)$$

where λ_a is the audio stream weight, and λ_e is the EMA stream weight. In these experiments, the weights were adjusted experimentally to 0.7 and 0.3 values, respectively. The selected weights were obtained by maximizing the accuracy on several experiments.

2.3.3. Late fusion

A disadvantage of the previously described fusion method is the assumption of there being a synchrony between the two streams. Late fusion was applied to enable asynchrony between the audio and EMA streams, in this study. In the late fusion method, two single HMM-based classifiers were used for the audio speech and EMA speech, respectively. For each test utterance (i.e., isolated phone), the two classifiers provided an output list, which included all the phone hypotheses along with their likelihoods. Following that, all the separate mono-modal hypotheses were combined into bi-modal hypotheses using the weighted likelihoods, as it is given by,

$$logP_{AE}(h) = \lambda_a logP_A(h|\Omega_A) + \lambda_e logP_E(h|\Omega_E) \quad (4)$$

where $logP_{AE}(h)$ is the score of the combined bi-modal hypothesis h, $logP_A(h|\Omega_A)$ the score of the h provided by the audio classifier, and $logP_E(h|\Omega_E)$ the score of the h provided by the EMA classifier. λ_a and λ_e are the stream weights with the same constraints applied in multistream HMM fusion.

The procedure described here finally resulted in a combined N-best list, in which the top hypothesis was selected as the correct bi-modal output. A similar method was also introduced in [1].

3. RESULTS

3.1. Phoneme recognition in clean environment

Figure 2 shows the results obtained for vowel-, consonant-, and phoneme classification using EMA and audio parameters based on multistream HMM decision fusion. It is observed that EMA

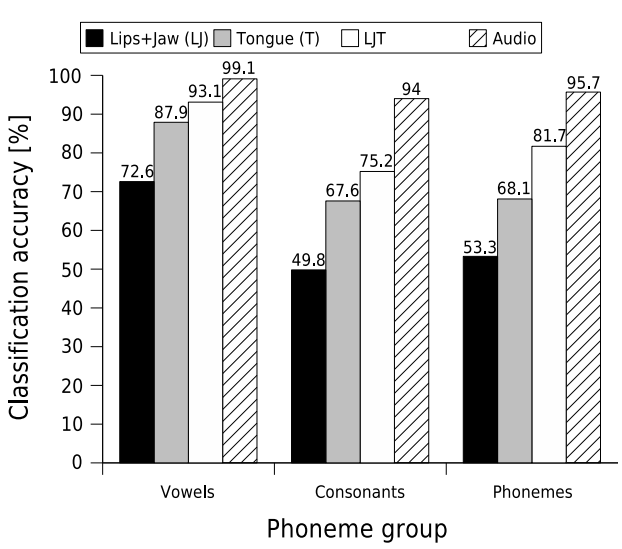

Fig. 2: Phoneme recognition results using EMA parameters.

can capture the speech information with very high accuracy. In addition, the results show that the EMA T parameters can capture the speech information better than the EMA LJ parameters. Integrating the LJ and T parameters leads to higher accuracy than that for LJ or tongue separately. Also, it can be observed that the vowel classification accuracy when using EMA parameters is 93.1% compared with the 99.1% classification accuracy when audio parameters are used. In the case of consonant and phoneme recognition, however, larger differences are obtained. However, since EMA cannot capture the voicing, a higher number of confusions appear between voiced and unvoiced consonants articulated in the same place (e.g., /p/ and /b/, /t/ and /d/, etc.).

3.2. Phoneme recognition in noisy environments

In these experiments, simulated noisy data on several signal-to-noise ratio (SNR) levels were fused with the EMA parameters, and were tested using multi-condition HMMs (i.e., HMMs trained using EMA data and noisy audio data on different SNR levels). The first stream consisted of 12 MFCC, 12ΔMFCC, and 12 $\Delta\Delta$MFCC parameters. The second stream consisted of 12 EMA PCA parameters, along with the first and second derivatives.

Figure 3 shows the comparison of the three fusion methods where white noise was used. In the case of SNR with a -10 dB level, the accuracy when using the feature fusion was 79.1%, when using the multistream HMM fusion it was 81.3%, and when using late fusion it was 83.4%. In the case of clean speech, the accuracy when using the feature fusion was 85.67%, when using the multistream HMM fusion it was 92.1%, and when using the late fusion it was 94.1%. It was seen that the highest accuracy was achieved when late fusion was used. In contrast, the lowest performance was obtained when concatenative fea-

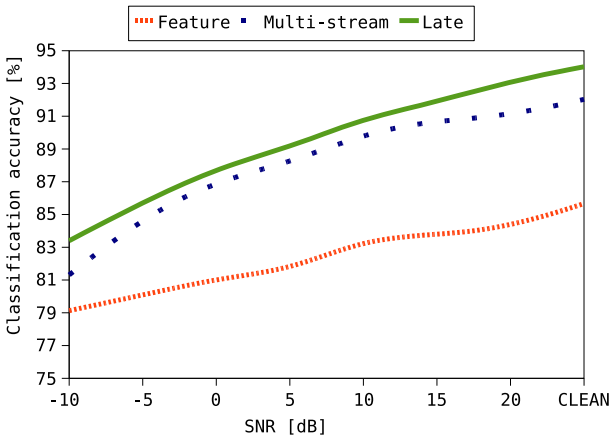

Fig. 3: Comparison between the three fusion methods (white noise).

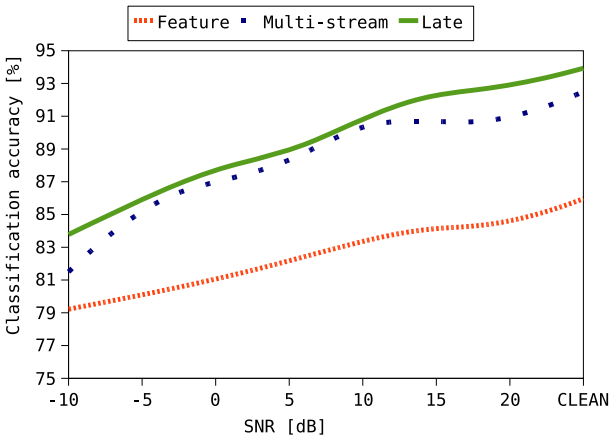

Fig. 5: Comparison between the three fusion methods (factory noise).

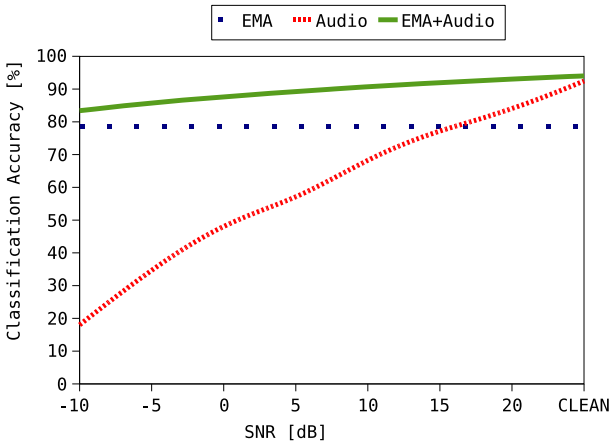

Fig. 4: Recognition in noisy environment using audio and audio-EMA parameters (white noise).

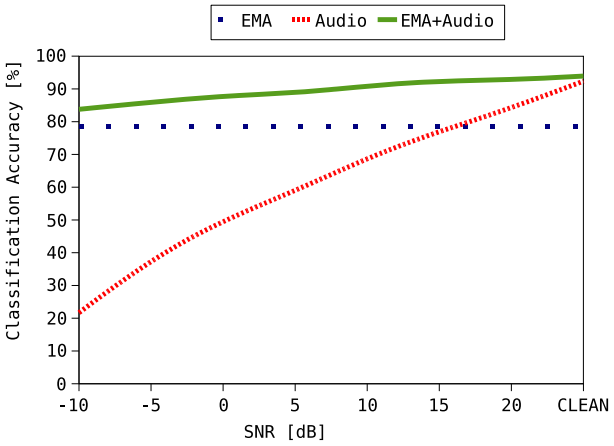

Fig. 6: Recognition in noisy environment using audio and audio-EMA parameters (factory noise).

ture fusion was used. A possible reason could be that feature fusion cannot capture the reliability of each stream. Another possible reason for lower accuracy might be inadequate modeling due to the higher dimension of the feature vectors, and also due to the limited training data.

In order to assess the possible contribution of the EMA parameters to the recognizer's robustness against noise, experiments were conducted using fused EMA-audio data, based on late fusion. The obtained results were compared with the results of experiments where noisy audio data were tested using multi-condition noisy audio HMMs.

Figure 4 shows the results obtained in the case of white noise. It is observed from the figure that the integration of EMA parameters significantly increases the accuracy compared to noisy audio speech. It is also observed, that the accuracy when the EMA features are fused with audio speech is superior to both cases where EMA and audio speech are used in isolation. In the case of SNR with -10 dB level, the accuracy when EMA parameters are also fused is 83.4%, as compared to an 18.0% accuracy when only acoustic parameters are used. In the case of clean speech, the accuracy when using fused EMA-audio speech is 94.1%, and 92.5% when only audio parameters are used. When using the EMA parameters alone, the accuracy is 78.7% for all SNR levels.

Another set of experiments were conducted using factory noise. Figure 5 shows the comparison of the three fusion methods. The results obtained were very similar to the case where white noise was used. In the case of SNR with -10 dB level, the accuracy when using the feature fusion was 79.22%, when using the multistream HMM fusion it was 81.5%, and when late fusion was used it was 83.9%. When clean speech was used, 86% accuracy was achieved when the feature fusion was used, 92.5% when multistream HMM fusion was used, and 93.9% when late fusion was used. As it is also shown in the case of using factory

noise, the highest accuracy was obtained when late fusion was used, and the lowest accuracy was obtained when feature fusion was used.

Figure 6 shows the results obtained in the case of using factory noise. Also in this case, the integration of EMA parameters significantly increased the accuracy. At -10 dB SNR level, the accuracy when using audio features alone was 21.7%. However, when the EMA parameters were integrated the accuracy rose to 83.9%. For clean audio speech, the accuracy when using only the audio parameters was 92.4% compared to 93.9% accuracy when using fused EMA-audio parameters.

4. CONCLUSIONS

In this study, automatic recognition based on Electro-Magnetic Articulography was introduced. Using the movements of the tongue, lips, and jaw features, HMMs were created and automatic phoneme recognition experiments were conducted, obtaining 78.7% phoneme classification accuracy. In particular, the authors were interested in demonstrating the contribution of the EMA parameters to robust speech recognition. Using three fusion methods, noisy speech was fused with the EMA parameters and tested with multi-conditional HMMs. The results showed that when the EMA parameters were fused with the audio parameters, the accuracy in noisy environments significantly increased. Specifically, in the case of SNR with a -10 dB level, 64% absolute increase in accuracy was obtained.

5. REFERENCES

[1] G. Potamianos, G. Gravier, A. Garg, A.W. Senior. Cooley, and J. W. Tukey, "Recent advances in the automatic recognition of audiovisual speech," *in Proc. of the IEEE*, vol. 91, Issue 9, pp. 1306–1326, 2003.

[2] R. O.Cornett, "Cued speech," *American Annals of the Deaf*, vol. 112, pp. 3–13, 1967.

[3] P. Heracleous, N. Abboutabit, and D. Beautemps, "Lip shape and hand position fusion for automatic vowel recognition in cued speech for french," *in IEEE Signal Processing Letters*, vol. 16, pp. 339–342, 2009.

[4] K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano, "Evaluation of speaking-aid system with voice conversion for laryngectomees toward its use in practical environments," *in Proc. of Intespeech*, pp. 2209–2212, 2008.

[5] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Unvoiced speech recognition using tissue-conductive acoustic sensor," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[6] T. Toda and K. Shikano, "Nam-to-speech conversion with gaussian mixture models," *in Proc. of Interspeech*, pp. 1957–1960, 2005.

[7] V. A. Tran, G. Bailly, H. Loevenbruck, and C. Jutten, "Improvement to a nam captured whisper-to-speech system," *in Proc. of Interspeech*, pp. 1465–1468, 2008.

[8] S.C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," *in Proc. of ICSLP*, pp. 573–576, 2006.

[9] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," *in Proc. of Interspeech*, pp. 2032–2035, 2008.

[10] E. Uraga and T. Hain, "Automatic recognition experiments with articulatory data," *in Proc. of ICSLP*, pp. 353–356, 2006.

[11] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech system for patients following laryngectomy," *Medical Engineering & Physics*, vol. 30, Issue 4, pp. 419–425, 2008.

[12] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," *in Proc. of ICSLP 2000*, pp. 145–148, 2000.

[13] J. S. Perkell, M. M. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabiet, , and M. T. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America*, vol. 92, pp. 3078–3096, 1992.

[14] Y. Tarabalka, P. Badin, F. Elisei, and G. Bailly, "Can you read tongue movements? evaluation of the contribution of tongue display to speech understanding," *in 1ère Conférence internationale sur l'accessibilité et les systèmes de suppléance aux personnes en situation de handicaps (ASSISTH'2007) (N. Vigouroux, P. Gorce, and J.-L. Nespoulous, Eds.)*, pp. 187–193, 2007.

[15] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," *in Proc. of ICSLP*, pp. 426–429, 1996.