EXPERIMENTS OF SPEECH RECOGNITION IN A NOISY AND REVERBERANT ENVIRONMENT USING A MICROPHONE ARRAY AND HMM ADAPTATION

D. Giuliani, M. Omologo and P. Svaizer

IRST – Istituto per la Ricerca Scientica e Tecnologica
I–38050 Povo,Trento, ITALY

ABSTRACT

The use of a microphone array for hands-free continuous speech recognition in noisy and reverberant environment is investigated. An array of four omnidirectional microphones is placed at 1.5 m distance from the talker; given the array signals, a Time Delay Compensation (TDC) module provides a beamformed signal, that is shown effective as input to a Hidden Markov Model (HMM) based recognizer. Given a small amount of sentences collected from a new speaker in a real environment, HMM adaptation further improves recognition rate. These results are confirmed both by experiments conducted in a noisy office environment and by simulations. In the latter case, different SNR and reverberation conditions were recreated by using the image method to reproduce synthetic array microphone signals.

1. INTRODUCTION

This work presents recent results of a project that is under way at IRST laboratories, for the development of a hands-free microphone-array based dictation system. The system uses a Continuous Density HMM-based speech recognizer [1] trained with a large speech corpus acquired in a quiet room using a high quality close-talk microphone [2]. The four-microphone array acquisition system derives from that developed for acoustic surveillance purposes, as described in [3].

A previous work [4] reported on some recognition experiments conducted in a real noisy office environment and showed performance improvement due to the use of the microphone array. The remaining mismatch between training and test conditions was addressed using different compensation/enhancement techniques. Results demonstrated a further considerable improvement using phone HMM adaptation. This paper aims at predicting system behaviour under different environmental conditions. A block diagram of the experimental set-up is shown in Figure 1. The new results are obtained using a simulation approach that allows, in principle, to recreate a wide variety of noisy and reverberant conditions and includes geometrical informations such as room size and position of talker, noise source, and microphone array.

2. SYSTEM DESCRIPTION

2.1. Linear Microphone Array

The use of a microphone array for hands-free speech recognition [5] relies on the possibility of obtaining a signal of improved quality, compared to the one recorded by a single microphone. A microphone array system allows to emphasize the talker message, as well as to reduce noise and reverberation components, in a way that can be considered "independent" of the talker position.

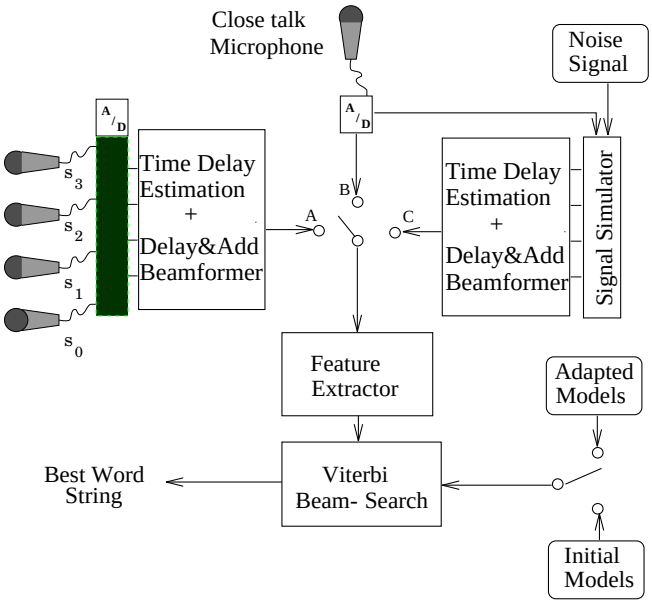

Figure 1: Block system diagram that includes three experimental set-up as well as the possible use of adapted HMMs. In particular, the switch on A corresponds to real data experiments, while the switch on C corresponds to simulations.

Let us assume that a talker generates an acoustic event $s(t)$ that is acquired by microphones $0,.....,(M-1)$ as signals $s_0(t),....,s_{M-1}(t)$. Signals sampled by acoustic sensors i and k are characterized by a relative delay δ_{ik} of the direct wavefront arrival. Time delay estimation is a critical issue un-

der noisy and reverberant conditions: in this work we adopted a CrosspowerSpectrum Phase (CSP) technique, that was shown to be effective for acoustic event detection and location [3]. Once each relative delay $\hat{\delta}_{0k}$ of direct wavefront arrival between microphone 0 and k has been estimated, the simplest technique to reconstruct an enhanced version $\hat{s}(t)$ of the acoustic message is based on a Time Delay Compensation (delay and sum beamformer):

$$\hat{s}(t) = \frac{1}{M} \sum_{k=0}^{M-1} s_k(t + \hat{\delta}_{0k}). \quad (1)$$

With a linear array of few microphones only a moderate directivity of acquisition over a restricted bandwidth can be achieved. Therefore, the corresponding enhancement capabilities may be reduced by the presence of noise propagation and reflected wavefronts in the steering direction. In order to face with these drawbacks and obtain a high spatial selectivity, more sophisticated acquisition systems should be employed, such as 2D microphone arrays with multiple band nested sensors [6].

2.2. Acoustic Feature Extraction

The input to the Feature Extractor (FE) corresponds to the digital version of the close-talk microphone in the case of the baseline system, and to the TDC processing output (1) when the microphone array is used for acquisition.

The FE input signal is preemphasized and blocked into frames of 20 ms duration. For each frame, 8 Mel scaled Cepstral Coefficients (MCCs) and the log-energy are extracted. MCCs are normalized by subtracting the MCC means computed on the whole utterance. The log-energy is also normalized with respect to the maximum value in the sentence. The resulting MCCs and the normalized log-energy, together with their first and second order derivatives, are arranged into a single observation vector of 27 components.

2.3. HMM-based Recognition System

A set of 34 context independent acoustic-phonetic speech units is modeled with left-to-right CDHMMs. Output distribution probabilities are modeled by means of mixtures having 16 Gaussian components with diagonal covariance matrices. Model training was accomplished by using the italian database APASCI [2]. The training set consisted of 2140 sentences uttered by 100 speakers (50 males and 50 females).

2.4. HMM Adaptation

In this work, an adaptation technique, based on Maximum a Posteriori (MAP) estimation [7] of model parameters, is used for HMM adaptation both to the new channel and to the speaker.

Only the Gaussian means are adapted while all the other parameters of the initial models are left unchanged. Speaker-independent models are used both as initial models and for setting prior parameters (e.g. each Gaussian mean vector of the initial models is used as the mean of an a priori Gaussian distribution). Let $\boldsymbol{m}_k$ be the mean vector of the k-th component of a mixture Gaussian distribution of an initial model. Under some assumptions [7] the MAP re-estimate of the k-th Gaussian mean can be formulated as:

$$\hat{\boldsymbol{m}}_k = \frac{c_k}{\tau_k + c_k} \boldsymbol{m'}_k + \frac{\tau_k}{\tau_k + c_k} \boldsymbol{m}_k \quad (2)$$

where c_k denotes the count observed for the k-th Gaussian component after an iteration of a conventional training algorithm exploiting the adaptation data and $\boldsymbol{m'}_k$ is the corresponding Maximum Likelihood estimate of the k-th Gaussian mean. τ_k is a parameter controlling the relative weight of the prior knowledge and the adaptation data. In this work, the same value of τ_k was adopted for all the Gaussian components.

	CtMic	*Mic0*	*Array*
Baseline	19.0	68.5	49.7
Ada	14.5	34.2	26.2

Table 1: Average WER(%) measured on the 240 sentences of the four speaker test sets.

3. EXPERIMENTS AND RESULTS

3.1. Multichannel Speech Corpus

A multichannel corpus was collected in an office environment to measure real system performance as well as to make a comparison with that predicted by simulation. Due to the characteristics of the room, recordings included a small amount of reverberation (reverberation time $T_{60} = 0.25s$), as well as coherent noise due to secondary sources (e.g. computers, air conditioning, etc). Multichannel recording of each utterance was accomplished by using both a close-talk cardioid microphone (following called *CtMic*) and a linear microphone array (following called *Array*) situated in front of the speaker at 150 cm distance. The array consisted of four microphones: distance between microphones was 30 cm. For comparison purposes, the array microphone *Mic0* was also used as an independent acquisition channel.

Eighty sentences were uttered by four speakers (2 males and 2 females). For each speaker, a development set and a test set were defined, that consisted in 20 sentences and 60 sentences, respectively. Each development set was used to adapt phone HMMs of the given speaker. Each test set included 789 words (13492 phone-like units) and was characterized by a word dictionary size equal to 343. Word Error Rate (WER) was measured given a Word Loop (WL) grammar having a single state and a self-loop per word; the resulting perplexity was equal to the dictionary size. Acquisitions were carried out synchronously for all the input channels at 16kHz sampling frequency, with 16 bit accuracy. Signal to Noise Ratio (SNR), measured as ratio between speech energy and noise energy, according to a speech-noise classification,

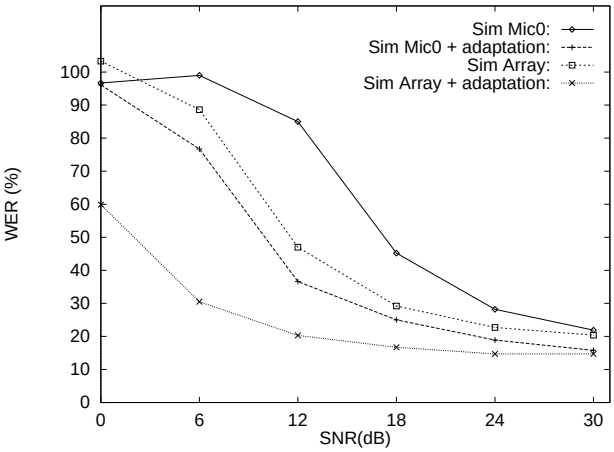

Figure 2: Word Error Rate (WER) for different SNR levels obtained adding *real* noise to microphone signals. A close talk microphone signal was used to derive simulated signals for a four microphone array (*Array*) and for a single remote microphone (*Mic0*). Experiments were carried out both with speaker-independent models and with adapted ones.

was estimated as 25 dB for *CtMic* and 18 dB for *Mic0* material. It is worth noting that an automatic segmentation and labeling system was preliminarily applied to the *CtMic* signals. A manual checking was then conducted to improve phone boundary alignment at the beginning and at the end of each sentence, that is where very small segmentation errors could cause a considerable fluctuation of the resulting SNR. Finally, this segmentation was used to derive the corresponding phone boundaries for each array signal.

3.2. Simulation Approach

Speech acquisition under different controlled environmental situations is problematic, especially if various conditions of reverberation need to be reproduced. For this reason only a typical office situation was considered for the acquisition of real signals at distance, with an array of microphones. Different conditions were then recreated by means of simulations starting from data simultaneously acquired by *CtMic* microphone, and therefore virtually free of noise and reverberation. In order to reproduce the effect of various amounts of reverberation, *CtMic* signals were convolved with different room acoustic impulse responses from the speaker to each microphone. These impulse responses were derived by means of the "image method" [8] that assumes that acoustic wavefronts propagating in an enclosure behave as geometrical rays obeying the reflection law. This condition is fulfilled in the frequency range in which the dimensions of the walls are large compared with the acoustic wavelength. The image method considers reflected rays as originated from mirror images of the source, with reduced power according to the reflection coefficients of the walls. Multiple reflection of a ray is represented by image sources of higher order. Starting from a geometrical model of the acquisition room (5.5m by 3.6m by 3.5m), including talker and microphone coordinates, reverberation times in the range from 0.1 s to 1.0 s were obtained assuming proper values for the reflectivity of the walls. The desired SNR was generated by adding real noise (acquired very close to a computer fan of the mentioned office) of appropriate power. The competitive noise source was supposed concentrated in a single point in the simulated room and propagation under reverberant condition was reproduced to derive the noise component at each microphone of the array.

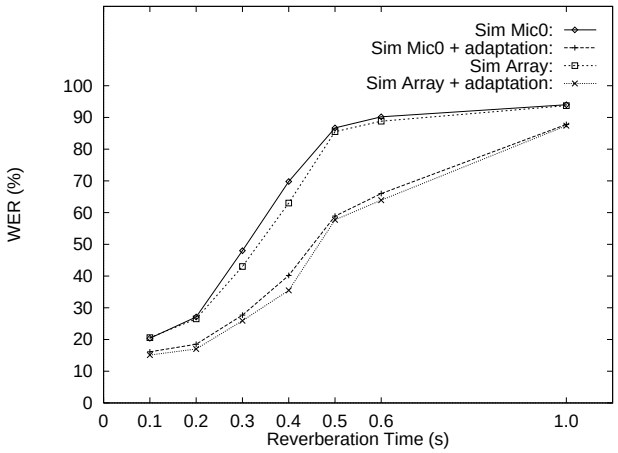

Figure 3: Word Error Rate (WER) from simulation experiments of only reverberant environment with different reverberation times T_{60}.

3.3. Real Data Experiments

Experiments on real data [4] showed performance reported in Table 1. The baseline system provided 19% and 68.5% WER when signals were acquired with the *CtMic* and *Mic0* of the array, respectively. The successive use of adapted HMMs improved system performance in both cases, providing 14.5% WER and 34.2% WER, respectively. Further improvement was obtained in the case of microphone array input leading to 49.7% WER and 26.2% WER, without and with HMM adaptation respectively.

3.4. Simulated Data Experiments

New system performance obtained through simulation experiments are reported in Figures 2, 3, 4, 5. Results refer to input simulation of a single remote microphone (*Sim Mic0*) and of the whole array (*Sim Array*).

The first experiment was conducted producing input signals that simulate the presence of a real noise source only. As shown in Figure 2, one can note that a clear advantage can be expected using the microphone array and the HMM adaptation for all the SNRs.

Figure 3 refers to the simulation of different reverberant conditions without any noise source. In this case, results show a

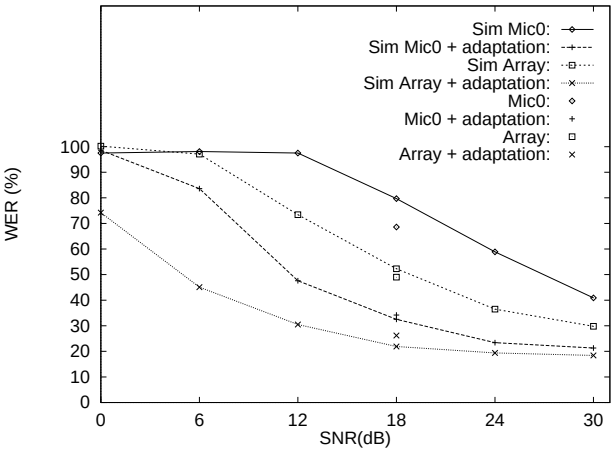

Figure 4: Word Error Rate (WER) from simulation experiments of both noisy and reverberant environment, given a reverberation time $T_{60} = 0.2s$.

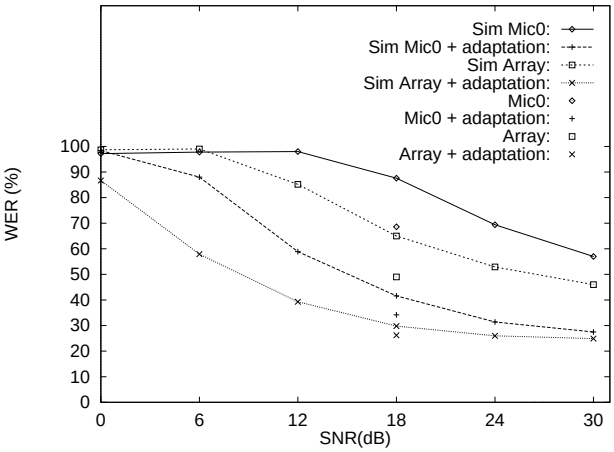

Figure 5: Word Error Rate (WER) from simulation experiments of both noisy and reverberant environment, given a reverberation time $T_{60} = 0.3s$.

clear improvement due to the HMM adaptation, while TDC provides a slight further error reduction. This behaviour can be explained by considering that the input signal of each array microphone includes reverberation components consisting in many replicas of the original signal coming from different directions, as pointed out in Section 2.1. Under those unreal conditions, the use of a linear microphone array with the proposed TDC processing is not adequate to reduce mismatching (between training and testing conditions), beyond what the HMM adaptation can do. A two-dimensional array consisting of a higher number of microphones is expected to improve both spatial selectivity and the quality of the signal reconstructed through the TDC processing [6].

Figures 4, 5 show results of simulation conducted assuming both the presence of a noise source and realistic reverberant room conditions. Note that real results are reported as single points of column at 18 dB SNR and that the multichannel corpus had been collected in a room characterized by an estimated reverberation time $T_{60} = 0.25s$. Recognition results obtained with simulated signals agree with results of experiments with real data as shown in Figures 4, 5. Performance improvement can be distinguished in the different contributions due to each added module, that is using or not the array and adapted HMMs. As seen in Figure 2, also in these cases the simple combination of TDC processing and an adapted HMM recognizer gives significant benefits for all SNRs.

4. CONCLUSIONS

This paper showed hands-free speech recognition performance that can be obtained including a microphone array and a HMM adaptation module in a HMM-based continuous speech recognizer. Recognition experiments were carried out with both real and simulated data. The simulation method here adopted turns out to be a precious tool for predicting performance capabilities of next versions of the recognizer, under a wide variety of environmental characteristics. Next work will focus on the verification of this experimental approach for different geometrical models and, in particular, using 2-D microphone arrays.

5. REFERENCES

1. S. Young, "Large Vocabulary Continuous Speech Recognition: a Review", *IEEE Workshop on ASR 1995*, Snowbird, December 1995, pp. 3–28.
2. B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo, "Speaker Independent Continuous Speech Recognition using an Acoustic-Phonetic Italian Corpus", *Proc. ICSLP*, Yokohama, September 1994, Vol. 3, pp. 1391–1394.
3. M. Omologo, P. Svaizer, "Acoustic Source Location in Noisy and Reverberant Environment using CSP Analysis", *Proc. ICASSP*, Atlanta 1996.
4. D. Giuliani, M. Omologo, P. Svaizer, "Robust Continuous Speech Recognition using a Microphone Array", *Proc. Eurospeech*, Madrid, September 1995, pp. 2021–2024.
5. C. Che, Q. Lin, J. Pearson, B. de Vries, J. Flanagan, "Microphone Arrays and Neural Networks for Robust Speech Recognition", *ARPA Workshop on Human language Technology*, NJ, March 1994, pp. 342–348.
6. J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, M.M. Sondhi, "Autodirective Microphone Systems", *ACUSTICA*, vol. 73, 1991.
7. J.-L. Gauvain, C.-H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291–299, 1994.
8. J.B. Allen, D.A. Berkley, "Image Method for efficiently simulating small-room acoustics", *Journ. of Acoust. Soc. Amer.*, vol. JASA 65(4), April 1979, pp. 943-950.