USE OF DIFFERENT MICROPHONE ARRAY CONFIGURATIONS FOR HANDS-FREE SPEECH RECOGNITION IN NOISY AND REVERBERANT ENVIRONMENT

Diego Giuliani, Marco Matassoni, Maurizio Omologo, Piergiorgio Svaizer

IRST-Istituto per la Ricerca Scientifica e Tecnologica
38050 Povo di Trento, Italy. Tel. +39 461 314563, FAX: +39 461 314591, E-mail:omologo@itc.it

10.21437/Eurospeech.1997-135

ABSTRACT

In this work hands-free continuous speech recognition based on microphone arrays is investigated. A set of experiments was carried out using arrays having different numbers of omnidirectional microphones as well as different configurations. Both real and simulated array signals, generated by means of the image method, were used. An enhanced input to a recognizer based on Hidden Markov Models was obtained by a time delay compensation module providing a beamformed signal. HMM adaptation was used to realign the recognizer acoustic modeling to the given acoustic condition.

1. INTRODUCTION

Hands-free continuous speech recognition represents a challenging scenario: many experimental activities [1, 2, 3, 4, 5, 6] have been recently devoted to the enhancement of the speech signal and to the compensation of the acoustic mismatch between training and testing conditions.

This work concerns the use of a Continuous Density HMM (CDHMM) based speech recognizer trained with a large speech corpus of clean speech material. Starting from the signals acquired by means of a microphone array system, a Time Delay Compensation (TDC) module provides a beamformed input. Some recognition experiments were carried out in a noisy office environment and showed performance improvement due to the use of the microphone array with respect to the use of a single microphone. The mismatch between training conditions and testing conditions has been further reduced using a phone HMM adaptation technique.

In [7] we described experiments performed both on real environment data and on simulated data. As evidenced in that work, the simulation method is a precious tool for predicting performance capabilities of the recognizer, under a wide variety of noisy and reverberant conditions.

Both talker's position and number of microphones have a direct impact on system performance. The effect of talker's position was investigated in [8], while a preliminary study concerning the relationship between number of microphones and system performance was reported in [14]. The purpose of this work is to extend the latter study, considering other microphone placement solutions, based on linear array geometries and harmonic ones.

The paper is organized as follows. Section 2 provides an introduction to the microphone array processing. In Section 3 an overview of the recognition system is outlined. In Section 4 the multichannel speech corpus is presented together with the method for producing simulated data. Section 5 reports on recognition experimental results. Finally, in Section 6, some guidelines for future work are remarked.

2. MICROPHONE ARRAYS

The use of a microphone array for hands-free speech recognition relies on the possibility of obtaining a signal of improved quality, compared to the one recorded by a single microphone.

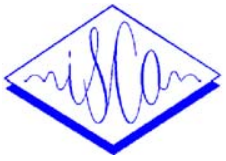
Let us assume that a talker produces a speech message $s(t)$ that is acquired by microphones $0, \ldots, (M-1)$ as signals $s_0(t), \ldots, s_{M-1}(t)$. Signals sampled by microphones i and k are characterized by a relative delay δ_{ik} of the direct wavefront arrival. Time delay estimation is a critical issue under 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 [9]. Once the 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 the following "delay and sum beamformer" TDC:

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

2.1. Linear Arrays

The beamwidth of the linear array beamformer is inversely proportional to the frequency, to the number of microphones and to the inter-microphone distance. Besides, if the array is characterized by a non adequately short distance between adjacent microphones, the so-called "spatial aliasing" effect occurs, i.e. secondary lobes (called *grating lobes*) of amplitude equal to the main lobe appear at the higher frequencies in the directivity pattern, along directions different from the desired one. Signals propagating from the directions of grating lobes cannot be discriminated from those propagating from the steering direction.

Figures 1d),e),f) show the directivity pattern at $2000Hz$, $4000Hz$, and $8000Hz$, when a linear array of five equispaced microphones (characterized by $5cm$ distance between adjacent microphones) is steered in the direction of $0°$. Note the grating lobes that are present in the latter case (with the given configuration, they appear for all the frequencies higher than $6500Hz$).



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

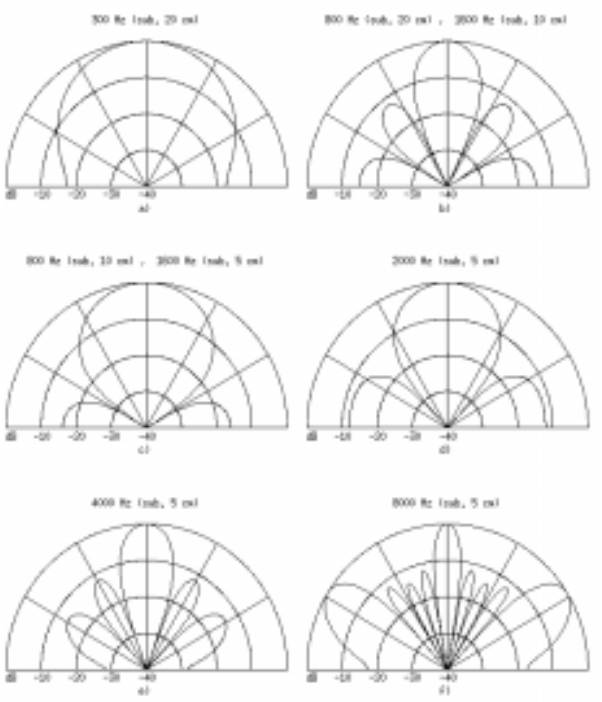


Figure 1: *Directivity patterns of five-microphone linear arrays, steered towards $\phi = 0°$. The patterns were evaluated at different frequencies between 300 Hz and 8000 Hz, for inter-sensor distances of 5 cm, 10 cm, and 20 cm.*

2.2. Harmonic Array

In order to achieve a better spatial selectivity and to reduce or eliminate the problem of spatial aliasing, either harmonically nested arrays or 2D microphones configurations [10] may be adopted. The harmonic array uses a distinct subarray for each frequency octave. In this way, the beamwidth remains unchanged across different frequency subbands, provided that the intersensor spacing is progressively halved from each octave to the higher one. Several octaves can be processed by harmonically nested subarrays, and the final output is obtained by subband recomposition. In this work, two harmonic arrays were considered both of them based on three subbands, namely: $[0, 800Hz], [800Hz, 1600Hz], [1600Hz, 8000Hz]$.

The harmonic array, shown in Figure 2d), is characterized by three nested subarrays, each consisting of 5 microphones. The intersensor spacing of the three subarrays are: 5 cm, 10 cm, and 20 cm. Figure 1 shows directivity patterns of the three subarrays that are nested to form the first harmonic array. Given this geometry and the above mentioned subband distribution, the two harmonic

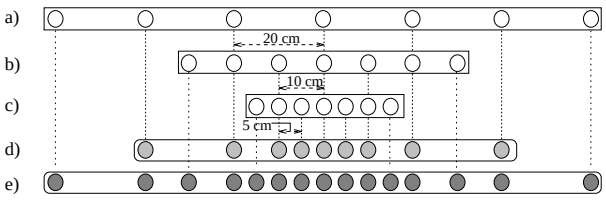


Figure 2: *Two harmonic arrays of 9 microphones (d) and 15 microphones (e), based on nesting three subarrays of different intersensor spacings (a,b,c).*

arrays are still characterized by spatial aliasing at higher frequencies, as shown in Figure 1f).

The harmonic array shown in Figure 2e) consists of 15 microphones placed in order to have 7 microphones in each of the three subarrays. Note that nine microphones of the former harmonic array belongs to the latter one as well.

3. SYSTEM DESCRIPTION

A block diagram of the whole recognition system is shown in Figure 3 (switch on A). The system consists of: a microphone array module that provides a beamformed output signal; a Feature Extraction (FE) module; a HMM-based recognizer that can operate either with speaker-independent HMM phone models or with "speaker and channel" adapted models. Figure 3 has also the purpose

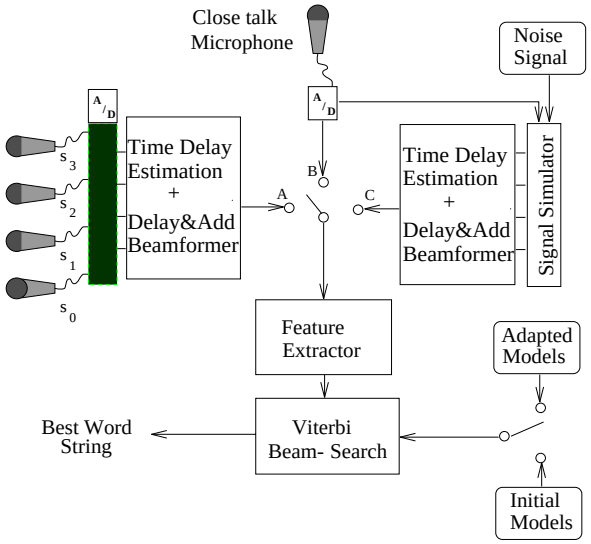


Figure 3: *Block diagram of the recognition system. Three input modalities are included: switch on A corresponds to real data experiments, switch on B to close-talk input, switch on C to simulated input.*

of highlighting two other ways of providing the input signal to the recognizer, namely: by using a close talk microphone (B) or by using a simulator of the microphone array processing (C). All these aspects will be detailed in the following.

3.1. Acoustic Feature Extraction

The input to the Feature Extractor (FE) is the signal acquired by a close-talk microphone in the case of the baseline system, and the output of the TDC processing (1) when the microphone array is used. 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.

3.2. HMM Recognizer

The recognition system is based on a set of 34 phone-like speech units. Each speech unit is modeled with left-to-right Continuous Density HMMs with output probability distributions represented by means of mixtures having 16 Gaussian components with diagonal covariance matrices. Model training was accomplished by using a phonetically rich italian corpus (APASCI) acquired in a quiet room by means of a high quality close-talk microphone [11].

Given the initial set of speaker independent HMMs, the mean vectors of the Gaussian mixture components are adapted according to a scheme based on Maximum a Posteriori estimation [12] and reported in [7].

4. MULTICHANNEL CORPUS

Speech data were collected in a large room ($10m \times 7m \times 3m$), characterized by a moderate amount of reverberation (reverberation time T_{60} was about $0.35s$) as well as by the presence of coherent noise due to some secondary sources (e.g. computers, air conditioning, etc). Eighty sentences were uttered by four speakers (2 males and 2 females) in a frontal position at 1.5 m distance from the array (F150). Multichannel recording of each utterance was accomplished by using both a close-talk cardioid microphone (*ClTalk*) and the linear microphone array (in the following called *8m-lin(10cm)*). Distance between the talker's mouth and the *ClTalk* microphone was approximately $15cm$. 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 at the microphones of the array was estimated as 30 dB.

Speech acquisition under different controlled environmental situations is problematic, especially if various conditions (i.e. noise, reverberation, talker position) need to be investigated. For this reason, some experiments were realized, that simulated speech propagation and acquisition (by each microphone of the array) in a room of the same size of that used for the real-data collection. Different situations were recreated, starting from data previously acquired by the *ClTalk* microphone, and therefore virtually free of noise and reverberation. In order to reproduce the effect of different array geometries and various amounts of noise and reverberation, each *ClTalk* signal was convoluted with room acoustic impulse responses from the speaker position to each microphone. These impulse responses were obtained by means of the "image method" [13] that assumes that acoustic wavefronts propagating in an enclosure behave as geometrical rays obeying the reflection law. This condition is only fulfilled in practice in the frequency range in which the dimensions of the walls are large compared with the acoustic wavelength. In the simulations we assumed a single competitive noise source concentrated where the noisiest source was present in the real-data collection (at approximately 3 m distance and at an angle of about 30°). The power of noise source was properly rescaled to obtain a desired average SNR for each speaker. Then noise propagation was simulated for each microphone of the array.

5. EXPERIMENTS AND RESULTS

For each speaker, an adaptation set and a test set were defined, that consisted in 20 sentences and 60 sentences, respectively. Each adaptation set was used to adapt the speaker-independent phone HMMs to acquisition channel, environmental condition, and speaker. Note that during the supervised adaptation procedure, alignment of the output of the TDC module against the initial models, for identification of silence and inter-word pauses, was first accomplished. The whole test set included 2316 words (14635 phone-like units) and was characterized by a word dictionary size equal to 343. Word Recognition Rate (WRR) was measured given a Word Loop (WL) grammar having a single state and a self-loop per word; hence, the resulting perplexity was equal to the dictionary size. Performance is represented as average WRR(%) measured on the test set consisting of the 240 sentences uttered by the four speaker.

	ClTalk	*Mic0*	*8m-lin(10cm)*
No Adapt	78.0	3.0	17.0
Adapt	83.9	31.5	65.5

Table 1: *Real environment experimental results. Performance is represented as average WRR(%) measured on the 240 sentences of the four speaker test sets. ClTalk, Mic0 and $8m-lin(10cm)$ indicate the three different front-end processing, used with or without phone HMM adaptation.*

5.1. Real Data Experiments

Table 1 provides system performance obtained with a talker at 1.5 m distance from the array. For comparison purposes, the first microphone of the array (*Mic0*) is also considered as an independent acquisition channel.

The joint use of the array processing and HMM adaptation always provides a definite improvement, with respect to the use of either a single microphone or the adaptation. As shown in [7], this result is confirmed for different talker positions and amount of noise and reverberation.

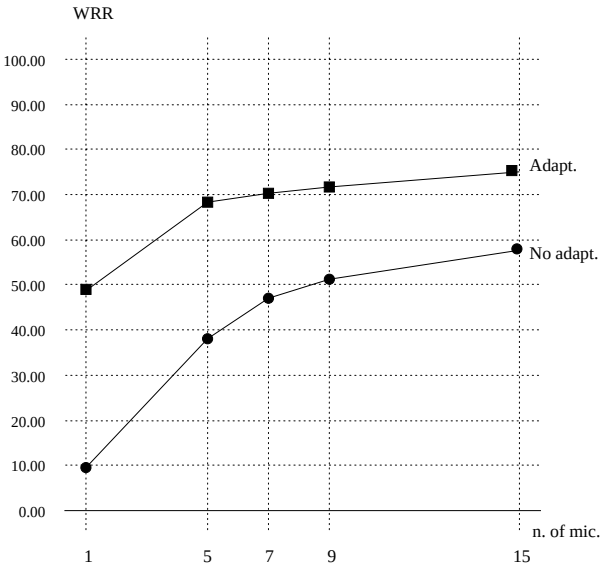


Figure 4: *Simulation results obtained using a different number of microphones, given a linear geometry and an environmental condition characterized by $T_{60} = 0.2s$ and an average SNR proximum to $13dB$.*

5.2. Simulated Data Experiments

Simulation experiments were conducted with $T_{60} = 0.2s$ and average SNR of $13dB$. Under these conditions, acquisitions based on linear array geometries as well as on harmonic ones were recreated. Figure 4 reports on system performance obtained changing the number of microphones. Results show a slight improvement increasing this number. However, under more hostile conditions the benefits of a higher array order would be more evident [7]. In the two cases of linear subarrays consisting of 5 and 7 microphones, system performance was evaluated also at different intersensor distances.

	5 cm	*10 cm*	*20 cm*	*30 cm*
No Adapt.	28.6	38.2	43.9	43.2
Adapt.	66.3	69.2	69.1	69.2

Table 2: *Simulation results obtained using a linear array of 5 microphones, given different intersensor distances ranging from 5 cm and 30 cm.*

	5 cm	*10 cm*	*20 cm*
No Adapt.	35.4	46.8	49.2
Adapt.	70.0	70.2	69.3

Table 3: *Simulation results obtained using a linear array of 7 microphones, given different intersensor distances.*

	9m-lin	*9m-harm*	*15m-lin*	*15m-harm*
No Adapt.	51.1	49.9	57.6	56.2
Adapt.	71.5	72.2	74.8	73.5

Table 4: *Simulation results obtained using linear arrays and harmonic arrays consisting of 9 and 15 microphones.*

The results, given in Tables 2 and 3, show a general improvement for larger spacing (e.g. 20 cm), when the recognizer is not adapted. On the other hand, the improvement is not observed anymore when the adaptation is applied.

Finally, Table 4 reports on performance obtained with the two harmonic arrays shown in Figure 2. In the case of 9 microphone array and HMM adaptation, the harmonic geometry slightly outperformed the linear one. However, the general trend does not show any convenience in using the harmonic array (even if it always performs better than each subarray nested in it).

6. FUTURE WORK

Harmonic arrays introduce a significant increase of complexity in the design and development of a hands-free recognition system. The results shown in this work are not much encouraging but have to be considered preliminary and to be confirmed under different environmental conditions, talker positions and noise source distribution in space.

Many other issues may be pursued in order to improve performance and robustness of the hands-free recognizer that is being studied. For instance, the use of other acoustic features as well as of post-processing techniques (e.g. adaptive post-filtering) applied to the beamformed signal could provide further improvement to the present system performance, in particular under very hostile conditions. Furthermore, the dependence of system behavior on discrepancies between talker position during training and testing (and the consequent influence of errors in the array steering) deserve to be investigated. Finally, new methods for phone HMM adaptation deserve to be investigated: a particular attention will be devoted to techniques that may be applied while the system is on-line and in an unsupervised manner.

7. REFERENCES

[1] 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.

[2] J.E. Adcock, Y. Gotoh, D.J. Mashao, H.F. Silverman, "Microphone-Array Speech Recognition via Incremental MAP Training" *Proc. ICASSP*, Atlanta 1996, pp. 897–900.

[3] Y. Grenier, "A microphone array for car environments", *Speech Communication*, vol. 12, 1993, pp. 25-39.

[4] R.M. Stern, F.H. Liu, Y. Ohshima, T. Sullivan, A. Acero, "Multiple Approaches to Robust Speech Recognition", *ARPA Workshop on Human language Technology*, NY, 1992, pp. 274–279.

[5] D. Van Compernolle, W. Ma, F. Xie, M. Van Diest, "Speech Recognition in Noisy Environments with the Aid of Microphone Arrays", *Speech Communication*, vol. 9, 1990, pp. 433-442.

[6] T. Yamada, S. Nakamura, K. Shikano, "Robust Speech Recognition with Speaker Localization by a Microphone Array", *Proc. of ICSLP*, Philadelphia, October 1996.

[7] D. Giuliani, M. Omologo, P. Svaizer, "Experiments of Speech Recognition in a Noisy and Reverberant Environment using a Microphone Array and HMM Adaptation", *Proc. of ICSLP*, Philadelphia, October 1996.

[8] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, "Microphone Array based Speech Recognition with different talker-array positions", *Proc. of ICASSP*, Munich, April 1997, pp.227-230.

[9] M. Omologo, P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location", *IEEE Trans. on Speech and Audio Processing*, May 1997, vol. 5, n. 3, pp. 288-292.

[10] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, M.M. Sondhi, "Autodirective Microphone Systems", *ACUSTICA*, vol. 73, 1991.

[11] 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.

[12] 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.

[13] 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.

[14] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, "Hands-free Speech Recognition in a Noisy and Reverberant Environment", *Proc. of the ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson (France), April 1997, pp. 195-198.