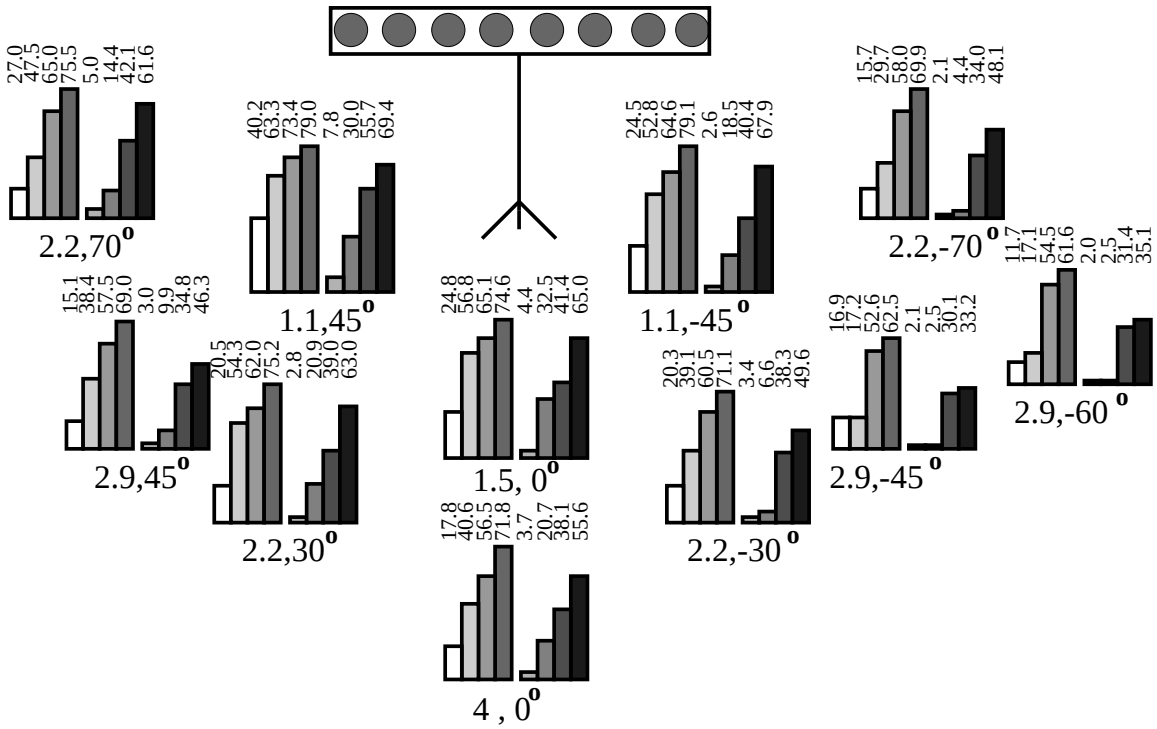

Figure 4. *Simulation experimental results. For each position (represented by its polar coordinates), two groups of WRRs are given, corresponding to the artificial environmental conditions $T_{60} = 0.2s$-$SNR = 13dB$ and $T_{60} = 0.3s$-$SNR = 10dB$. Each of the set of results consists of four WRRs that correspond, from left to right, to the following experimental conditions: Mic0, Array, Mic0 + Adaptation, Array + Adaptation.*

future work will be devoted to confirm the advantages of the hands-free recognition system, here presented, in the large vocabulary dictation system being developed at IRST labs.

However, from the results described above other important issues remain to be addressed. One is the use of array geometries alternative to the present linear one: both harmonic arrays and 2-D arrays represent promising solutions to be investigated. Another one is the study of new methods for phone HMM adaptation: a particular attention will be devoted to techniques that can be applied while the system is on-line and in an unsupervised manner. Also the use of other acoustic features, more robust than mel-based cepstral coefficients, and of adaptive post-filtering techniques (to apply to the beamformed signal) could provide further improvement to the present system performance. Finally, the dependence of system behavior on discrepancies between the talker position during training and during testing (and the influence of errors in the array steering) deserve to be investigated.

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] S. Young, "Large Vocabulary Continuous Speech Recognition: a Review", *IEEE Workshop on ASR 1995*, Snowbird, December 1995, pp. 3–28.

[9] M. Omologo, P. Svaizer, "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", *Proc. ICASSP*, Adelaide 1994, vol. 2, pp. 273-276.

[10] M. Omologo, P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location", to appear in *IEEE Trans. on Speech and Audio Processing.*

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

[12] 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.

[13] 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.

[14] 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.

energy and noise energy at the respective sources. Noise propagation was then simulated from the noise source to each microphone of the array, and the resulting signal was added to the simulated speech one.

Note that the talker positions were chosen not so close to the microphone array to determine a very near field condition, for which the adopted simulation method was not retained appropriate.

3.2. New multichannel Real-data Corpus

The multichannel corpus was collected in a large room of the mentioned size, and characterized by a moderate amount of reverberation (reverberation time $T_{60} = 0.35s$) 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. Multichannel recording of each utterance was accomplished by using both a close-talk cardioid microphone (*CtMic*) and the linear microphone array (in the following called *Array*). Distance between the mouth's talker and the *ClTalk* microphone was approximately $15cm$. For comparison purposes, the first microphone of the array (*Mic0*) was also used as an independent acquisition channel.

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 after having performed an automatic speech-noise classification as well as a manual check, was estimated as 30 dB for *CtMic*.

Once the real-data corpus was collected, three replicas of each utterance were obtained by means of a loudspeaker, that is by reproducing the original one acquired through the *ClTalk* microphone. Three source positions were used during this data collection, namely: in front of the array at 1.5 m distance and at 4 m distance; in a lateral position, at 2.9 m distance and $-30°$ angle. These positions represent a subset of those considered in the simulation experiments.

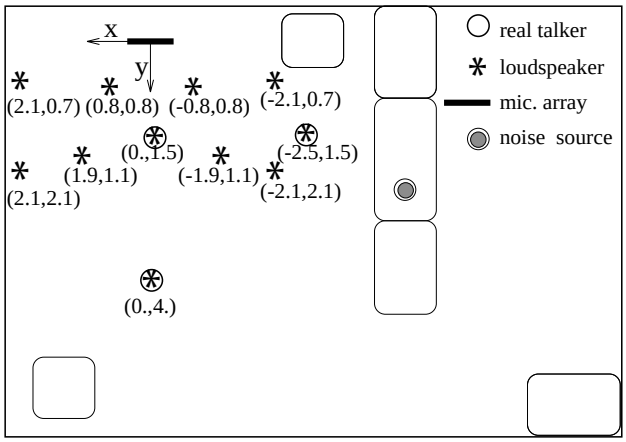

Figure 3. *Map of the experimental room (7m x 10m x 3m), showing the positions of talker, loudspeaker, microphone array, noise source and furniture.*

3.3. Recognition Task

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 then 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 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.

	ClTalk	*Mic0*	*Array*
RT-(0,1.5)	78.0	3.0	17.0
RT-(0,1.5)-Ada	83.9	31.5	65.5
LS-(0,1.5)	69.2	3.7	13.8
LS-(0,1.5)-Ada	78.8	40.0	64.4
LS-(0,4.0)	-	1.8	2.8
LS-(0,4.0)-Ada	-	13.6	36.6
LS-(-2.5,1.5)	-	1.5	2.8
LS-(-2.5,1.5)-Ada	-	14.1	43.8

Table 1. *Real environment experimental results. Performance is represented as average WRR(%) measured on the 240 sentences of the four speaker test sets, in the cases of Real Talkers and of LoudSpeaker positioned at three different distances from the array. ClTalk, Mic0 and Array indicate the three different front-end processing that were used both with and without phone HMM adaptation.*

3.4. Real Data Experiments

Table 1 provides performance for the Real Talkers (RT) and for the LoudSpeaker (LS) in the three given positions.

These results show a noticeable performance degradation due to the LS-array distance of the second and third positions and to the severe conditions under which the system was tested. Most of this degradation seems to be related to the LS-array distance, for which also the combination of TDC and HMM adaptation can provide limited performance improvement.

Note that a degradation (from 83.9% to 78.8% WRR) can be observed also passing from the RT case to the LS case for the position $(0, 1.5)$. This discrepancy was probably due to a varying distance between the talker and the *ClTalk* microphone (while that between LS and *ClTalk* was fixed to $15cm$) as well as to diffusion properties of the loudspeaker. This aspect deserves a further inspection, but does not prevent to make a consistent performance comparison at different LS positions.

3.5. Simulated Data Experiments

Two simulation sessions were performed, one with $\{T_{60} = 0.2s$ and $SNR = 13dB\}$ (SNR that was imposed at the speech and noise sources), the other with $\{T_{60} = 0.35s$ and $SNR = 10dB\}$. The latter one can be retained comparable to the condition of the real-data collection.

Figure 4 shows WRR with the speech source in the 11 positions, under the two simulated environmental conditions. Note that performance in positions $(-2.5, 1.5)$ and $(-2.1, 2.1)$, and in general on the left of the array, were influenced by the presence of the noise source at a low distance behind the speech source. In this case, simulation could not reproduce a situation equivalent to that of real-data collection, where more than one noise sources were distributed in space and one, in particular, behind the position $(0, 4)$.

Nevertheless, results are consistent and show the good system behaviour in some hostile conditions. In particular, the joint use of the array processing and HMM adaptation always provides a definite improvement, respect to the use of either one microphone of the array or of the adaptation only. In general, performance seems to be more dependent on the speech source-array distance than on the angle.

4. FUTURE WORK

The present system is being tested with an artificial recognition task, represented by a 343-word loop grammar. A

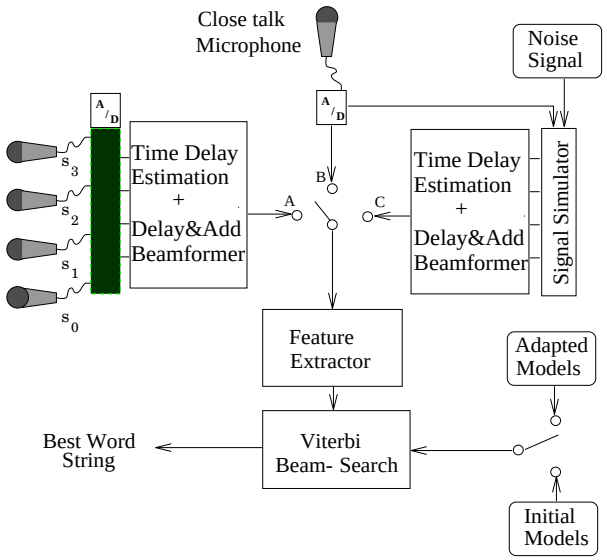

Figure 1. *Block system diagram that includes three experimental set-up as well as the possible use of adapted HMMs. In particular, the switch on* A *corresponds to real data experiments, while the switch on* C *corresponds to simulations.*

and to the inter-microphone distance). Besides, if the array is characterized by a non adequately low distance between adjacent microphones, the so-called "spatial aliasing" effect occurs, that is other lobes (called *grating lobes*) comparable to the main one appear 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. A way to reduce these effects is to use a higher number of microphones and different geometries, such as those of harmonic linear arrays or of 2D microphone arrays [11].

In the following, the analysis is limited to the use of a linear array of eight equispaced microphones, characterized by a 10 cm distance between adjacent microphones. Figure 2 shows the directivity pattern at $1000Hz$ and $2000Hz$, when this array is steered in the direction of $-30°$. Note that grating lobes are present, with this configuration, for frequencies higher than 2400 Hz.

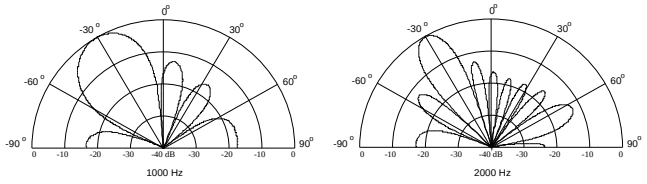

Figure 2. *Directivity patterns of the eight-microphone array steered towards $\phi = -30°$, evaluated at the frequencies f=1 kHz and f=2 kHz.*

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.

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 a phonetically rich italian corpus (APASCI) [12]. The training set consisted of 2140 sentences uttered by 100 speakers (50 males and 50 females).

2.4. HMM Adaptation

An adaptation technique, based on Maximum a Posteriori (MAP) estimation [13] of model parameters, is used for HMM adaptation both to the new acquisition 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 m_k be the mean vector of the k-th component of a mixture Gaussian distribution of an initial model. Under some assumptions, the MAP re-estimate of the k-th Gaussian mean can be formulated as:

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

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, m'_k is the corresponding Maximum Likelihood estimate of the k-th Gaussian mean, and $\tilde{m}_k$ is the prior mean vector. τ_k is a parameter controlling the relative weight of the prior knowledge and the adaptation data.

3. EXPERIMENTS AND RESULTS

3.1. Simulation Approach

Speech acquisition under different controlled environmental situations is problematic, especially if various conditions (noise, reverberation, talker position, etc) need to be investigated. For this reason, a simulation was realized of speech propagation and acquisition (by each microphone of the array) in a large room of the same size (10m by 7m by 3m) of that used for the real-data collection described in the next section. Different conditions were recreated, starting from data previously acquired by a close-talk (*CtMic*) microphone, and therefore virtually free of noise and reverberation. In order to reproduce the effect of different talker positions and various amounts of noise and reverberation, each *CtMic* signal was convolved with room acoustic impulse responses from the speaker to each microphone. These impulse responses were derived by means of the "image method" [14] 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.

Figure 3 shows the map of the room and 11 positions of the speech source. These positions were chosen in order to make a comparison of system performance both with different angles at a given distance from the array, and with different distances at the same angle. Simulations were realized assuming to have a single competitive noise source concentrated where the noisiest source was present in the real-data collection. For each utterance, the noise power was rescaled in order to have the same SNR between speech

MICROPHONE ARRAY BASED SPEECH RECOGNITION WITH DIFFERENT TALKER-ARRAY POSITIONS

Maurizio Omologo *Marco Matassoni* *Piergiorgio Svaizer* *Diego Giuliani*

IRST-Istituto per la Ricerca Scientifica e Tecnologica I-38050 Povo di 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 eight omnidirectional microphones was placed at different angles and distances from the talker. A time delay compensation module was used to provide a beamformed signal as input to a Hidden Markov Model (HMM) based recognizer. A phone HMM adaptation, based on a small amount of phonetically rich sentences, further improved the recognition rate obtained by applying only beamforming. These results were confirmed both by experiments conducted in a noisy and reverberant environment and by simulations. In the latter case, different conditions were recreated by using the image method to reproduce synthetic versions of the array microphone signals.

1. INTRODUCTION

In the last years, many experimental activities were devoted to investigate the use of microphone arrays for hands-free continuous speech recognition [1, 2, 3, 4, 5, 6].

The system under study at IRST laboratories [7] is based on a Continuous Density HMM - speech recognizer [8] trained with a large speech corpus acquired in a quiet room using a high quality close-talk microphone. In a previous work, a four-microphone array acquisition system was used to locate the talker [9] and reconstruct a beamformed signal, through a Time Delay Compensation (TDC) processing, that represented the input of the recognizer.

Some recognition experiments were conducted in a noisy office environment and showed performance improvement due to the use of the microphone array with respect to the use of one microphone. The mismatch between training and test conditions was further addressed using a phone HMM adaptation technique.

Both real environment and simulation experiments were described in [7]: from that work, the simulation method turned out to be a precious tool for predicting performance capabilities of the recognizer, under a wide variety of noisy and reverberant conditions. Results showed that the combination of the TDC processing and the adapted HMM recognizer gives significant benefits. Results evidenced also some limits of using only four microphones under reverberant conditions.

Another important aspect that remained to be addressed was the influence of the talker position on the system performance. In fact, the above mentioned results referred to a single talker position (in front of the array at 1.5 m distance).

This paper has the objective of extending the previous work evaluating system performance when an eight-microphone linear array is used and focusing the attention on the influence of the talker position on the recognition rate.

For this purpose, a new database has been collected in a noisy and reverberant large room and a new set of real-data experiments and simulations has been performed, as described in the following.

2. SYSTEM DESCRIPTION

A block diagram of the recognition system is shown in the Figure 1 (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 adapted ones. The Figure 1 has also the purpose of highlighting two other ways of providing the input signal to the recognizer, namely: using a close talk microphone (B) or using a simulator of the microphone array processing (C). All these aspects will be detailed in the following.

2.1. Linear Microphone Array

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. A microphone array system allows to emphasize the talker message, as well as to reduce noise and reverberation components, so that it could be used to make a system working "independently" of the talker position.

Let us assume that a talker produces a speech message $s(t)$ that is acquired by microphones $0, ..., (M-1)$ as signals $s_0(t)$,....,$s_{M-1}(t)$. 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 [10]. 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 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, even without errors in the delay estimates, only a moderate directivity of acquisition over a restricted bandwidth can be achieved.

If the array is steered in a direction that is different from that of the wavefront arrival, the spectrum of the array's output is distorted in a way depending of the beamformer transfer function. Typically, this distortion results in an attenuation of the high-frequency components of the source signal. This effect is mainly due to the fact that the beamwidth of the beamformer is inversely proportional to the frequency (as well as to the number of microphones,