EXPERIMENTS OF HANDS - FREE CONNECTED DIGIT RECOGNITION USING A MICROPHONE ARRAY

M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani
ITC - IRST Istituto per la Ricerca Scientifica e Tecnologica
38050 - Pante' di Povo
Trento - Italy

Abstract - A scenario concerning hands-free connected digit recognition in a noisy office environment is investigated. An array of six omnidirectional microphones and a corresponding time delay compensation module are used to provide a beamformed signal as input to a Hidden Markov Model (HMM) based recognizer. Phone HMM adaptation is used to further reduce the mismatch between training and test conditions.

Both real and simulated array signals, generated by means of the image method, were used to evaluate system performance. Real data were collected in a moderate noise environment with talkers at 1.5 m distance from the array. Results show that a digit accuracy over 95% can be achieved using the array and the HMM adaptation. This result has to be compared with 99.7% digit accuracy obtained by using a close-talk microphone instead of the array.

1 Introduction

Hands-free continuous speech recognition represents a challenging scenario: many experimental activities have been devoted to the enhancement of the speech signal and to the compensation of the acoustic mismatch between training and testing conditions. Microphone arrays [1] have been widely used for hands-free speech recognition [2, 3, 4, 5], thanks to the possibility of obtaining a signal of improved quality, compared to the one recorded by a single far microphone. Starting from the signals acquired by means of the microphone array system, a beamformed input is provided to a Continuous Density HMM based speech recognizer trained with clean speech. The performance improvement due to the use of the microphone array with respect to the use of a single microphone was addressed in [6, 7, 8, 9], where the mismatch between training and testing conditions was further reduced using a phone HMM adaptation technique. In those works, experiments performed both on real environment data and on simulated data also showed that the "image method" [10] is a precious tool for predicting performance capabilities of the

0-7803-3698-4/97/$10.00 © 1997 IEEE.

recognizer, under a wide variety of noisy and reverberant conditions. Furthermore, in those works other aspects were addressed such as: variabilities due to talker's position, optimization of the microphone array configuration.

The results provided in those works were obtained using a quite difficult recognition task, based on a word-loop grammar and a vocabulary consisting of 343 words. The purpose of this work is to extend some of the previous experimental activities to the task of connected digit recognition, that better represents a possible applicative context where this technology could be used in the next future.

2 System Description

A block diagram of the recognition system being studied is shown in Figure 1.

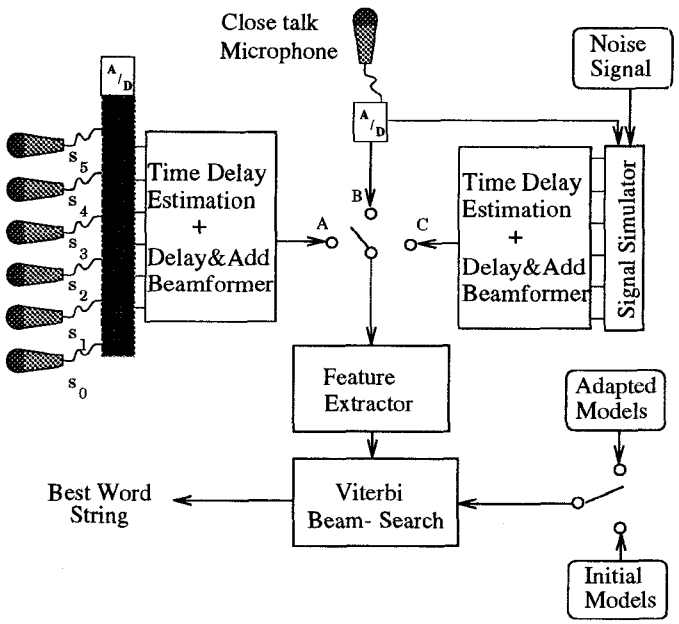

Figure 1: *Block diagram of the recognition system. Three input modalities are included: switch on A corresponds to real data experiments, switch on B to close-talk input, switch on* C *to simulated input.*

The hands-free recognition system (switch on A) consists of: a microphone array module that provides a beamformed output signal; a Feature Extraction module; a HMM-based recognizer that can operate either with clean HMM phone models or with adapted models. Figure 1 has also the purpose of highlighting two other ways of providing the input signal to the recognizer, namely: by using a close talk microphone (B) and by using a simulator of the microphone array processing (C).

2.1 Linear Microphone Array

The use of a microphone array [1] for hands-free speech recognition relies on the possibility of obtaining a signal of improved quality, compared to the one recorded by a single far microphone.

Let us assume that a talker produces a speech message that is acquired by the microphones of an array. Signals sampled by different microphones are characterized by a relative delay 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 [11]. Once the relative delays of direct wavefront arrival between microphones have been estimated, an enhanced version of the acoustic message is computed by applying Time Delay Compensation (delay and sum beamforming) as described in [7, 8, 9].

2.2 Speech Recognizer

The speech recognizer is based on 33 context-independent speech units represented by means of Continuous Density Hidden Markov Models (CDHMMs). A "silence" model has also been trained. Each output probability distribution is modeled with a mixture of 16 Gaussian densities having diagonal covariance matrices. Model training was performed exploiting the Italian continuous speech database APASCI. The training set consists of 2140 utterances acquired from 100 speakers (50 males and 50 females); recordings were performed in a quiet room with a close-talk microphone.

The acoustic front-end produces, for each frame (that is every 10 ms), 8 mel-scaled cepstral coefficients together with the frame log-energy. Each mel-scaled cepstral coefficient is normalized by subtracting its mean value computed utterance-by-utterance. The log-energy is also normalized by subtracting its maximum value computed on the whole current utterance. These acoustic parameters together with their first and second order time derivatives are arranged in a 27 observation vector.

2.3 HMM Adaptation

Given an initial set of CDHMMs, trained with clean speech in a speaker-independent mode, the mean vectors of the Gaussian components are adapted according to a scheme based on Maximum a Posteriori estimation [12]. Let $\mathbf{m}_k$ be the mean vector of the k-th component of a mixture of Gaussian densities. The adapted Gaussian mean $\hat{\mathbf{m}}_k$ is expressed as:

$$\hat{\mathbf{m}}_k = \lambda_k \mathbf{m'}_k + (1 - \lambda_k)\mathbf{m}_k \tag{1}$$

where $\mathbf{m'}_k$ is the Maximum Likelihood estimate of the k-th Gaussian mean obtained exploiting the available adaptation data, and λ_k is determined according to the relationship introduced in [12].

3 Multichannel speech corpus

Speech data were collected in an office ($5.5m \times 3.6m \times 3.5m$), characterized by a small amount of reverberation ($T_{60} \simeq 0.2s$) as well as by the presence of coherent noise due to some secondary sources (e.g. computers, air conditioning, etc). Figure 2 shows a map of the room and evidences the location of acoustic sources.

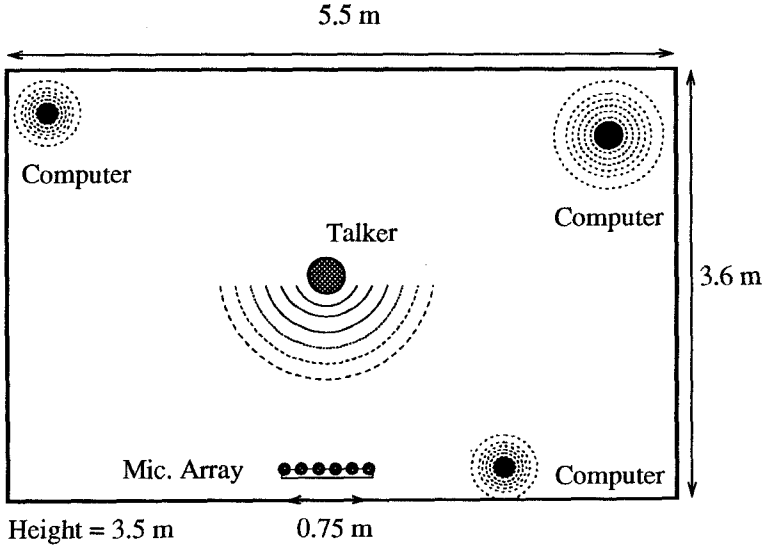

Figure 2: *Map of the experimental room (5.5m x 3.6m x 3.5m), showing the position of the talkers, of the microphone array and of computers.*

During a first recording session, 20 phonetically rich sentences (this set includes 210 words, among which there were no digits) were uttered by each of four speakers (2 males and 2 females) located in a frontal position, at 1.5 m distance from the array (F150). After two days, a second recording session was conducted in the same office, under similar environmental noise conditions, by collecting 50 connected digit strings (including 400 digit occurrences), from each of the four speakers (in the same position 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 *Array*). 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 (on the basis of an accurate manual speech-noise segmentation), was in the range between 12 dB and 18 dB. It is worth noting that SNR measured on close-talk microphone signals

was in the range between 24 and 33 dB.

4 Experiments and Results

For each speaker, a development set (first session) and a test set (second session) were defined: each development set was used to adapt clean phone HMMs to the acquisition channel, to the environmental condition as well as to the speaker. In the following experimental results obtained using both real data and simulated data are given. Performance is represented as Word Recognition Rate (WRR) measured on the whole test set (consisting of 1600 digit occurrences).

4.1 Real data

Table 1 reports on recognition performance obtained using speech data collected in the real environment. Using the *ClTalk* microphone as input, the system provides a 99.5% WRR and a 99.7% WRR in non-adapted and adapted mode, respectively.

	Baseline	*MAP Ad.*
ClTalk	99.5	99.7
Mic1	52.5	82.9
Array	74.3	95.4

Table 1: *Connected digit recognition results using real environment data.*

As shown in the Table, the use of a single far microphone *Mic1* (a microphone of the array) causes a noticeable performance degradation (from 99.5% to 52.5 % and from 99.7% to 82.9%, respectively).

N.ofAdapt.Utter.	*1*	*5*	*10*	*15*	*20*
ClTalk	99.5	99.6	99.7	99.7	99.7
Mic1	63.7	78.6	79.6	82.2	82.9
Array	78.6	92.8	94.0	95.0	95.4

Table 2: *System performance as a function of the adaptation material size.*

With the adoption of the array, the use of MAP adaptation allows a performance increase to 95.4% WRR. It is interesting to observe how this performance changes as a function of the adaptation material size, as reported in Table 2.

4.2 Simulation data

By simulation, different situations were recreated, starting from the signals acquired, during both recording sessions, by using the *ClTalk* microphone (close-talk microphone signals may be assumed virtually free of noise and reverberation).

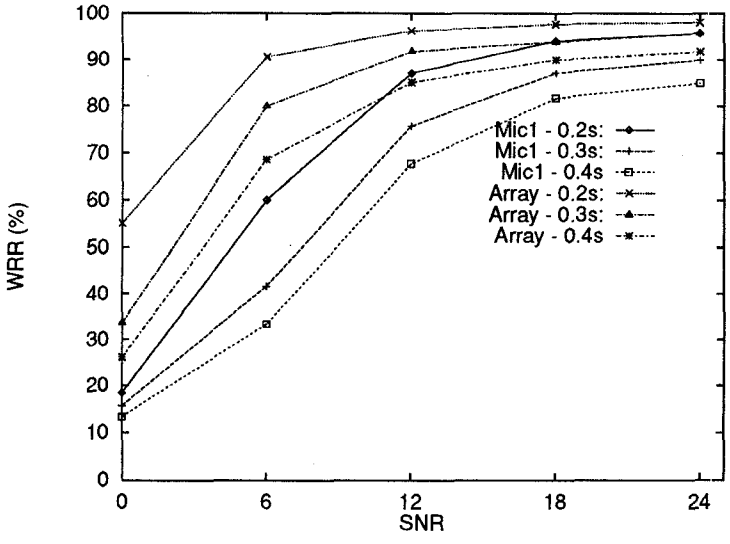

Figure 3: *Simulation results with MAP model adaptation as a function of SNR. Results were obtained for three different reverberation times, using either a single far microphone Mic1 or the Array as input to the recognizer.*

In order to reproduce the effects of a various amount of noise and reverberation, each *ClTalk* signal was convoluted with the room acoustic impulse responses from the speaker position to each microphone, obtained by means of the "image method" [10]. Furthermore, some competitive noise sources were located at the positions of the noisiest sources (computers) that were present in the office during real data recordings. The power of noise source was properly scaled to recreate the desired SNR between speech signal energy and noise energy at the respective sources. Then, noise propagation was simulated for each microphone of the array.

The results of all of the simulation experiments are given in Table 3: the behaviour of the system performance after MAP adaptation is also shown in Figure 3. From these results, one can note that there is a good performance matching between the real data experiments and the simulation based ones, conducted with a reverberation time around $0.2s$ and a SNR proximum to $12dB$. This fact confirms the usefulness of the simulation method, here adopted.

	$T_{60} = 0.2s$		$T_{60} = 0.3s$		$T_{60} = 0.4s$	
Mic1	*NoAd.*	*MAP*	*NoAd.*	*MAP*	*NoAd.*	*MAP*
24*dB*	83.8	95.8	70.8	90.0	62.7	85.0
18	74.7	94.1	60.6	87.1	52.2	81.7
12	50.0	87.1	37.4	75.8	30.4	67.7
6	15.2	60.0	13.6	41.6	13.6	33.3
0	15.4	18.4	15.6	15.6	16.0	13.3
Array	*NoAd.*	*MAP*	*NoAd.*	*MAP*	*NoAd.*	*MAP*
24*dB*	93.8	98.1	81.0	95.8	71.6	91.8
18	91.1	97.7	76.6	93.9	65.9	90.0
12	82.4	96.3	63.6	91.8	52.8	85.1
6	48.1	90.6	26.8	80.0	20.2	68.6
0	14.0	55.1	13.3	33.6	13.9	26.1

Table 3: *Results of simulation data experiments, with or without MAP adaptation, for different SNR levels and reverberation times.*

5 Conclusions and Future Work

Preliminary results have been provided that concern hands-free connected digit recognition experiments, conducted both on real data and on simulated data. The joint use of a microphone array input and of the MAP phone model adaptation allows to obtain a relevant performance improvement with respect to the use of a single far microphone input. It is worth noting that adaptation was performed by using data that were not collected in the test recording session.

Many issues remain to be addressed in order to extend the use of this technology to more complex situations. From simulations here described, one may conclude that more robust processing and adaptation techniques are needed for very adverse environmental conditions.

In the next future, adaptation based on Maximum Likelihood Linear Regression will be investigated. We will also focus our attention on adaptation techniques that can be applied while the system is on-line and in an unsupervised manner. A further activity will be the creation of a new experimental task that is the hands-free dictation of journal articles.

References

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

[2] 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.

[3] T.M. Sullivan and R.M. Stern, "Multi-Microphone Correlation-based Processing for Robust Speech Recognition", *Proc. ICASSP*, Minneapolis, April 1993, vol. 2, pp. 91–94.

[4] 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.

[5] M. Inoue, S. Nakamura, T. Yamada, K. Shikano, "Microphone Array Design Measures for Hands-free Speech Recognition", *Proc. of EUROSPEECH*, Rhodes, September 1997, pp. 331-334.

[6] 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, pp. 1329–1332.

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

[8] 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.

[9] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, "Use of Different Microphone Array Configurations for Hands-Free Speech Recognition in Noisy and Reverberant Environment", *Proc. EUROSPEECH*, Rhodes, September 1997, pp. 347-350.

[10] 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.

[11] 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.

[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.