EXPERIMENTS OF HMM ADAPTATION FOR HANDS-FREE CONNECTED DIGIT RECOGNITION

D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer

ITC-IRST - Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy
{giuliani,matasso,omologo,svaizer}@itc.it

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. Two different techniques of phone HMM adaptation have been considered, to reduce the mismatch between training and test conditions. Adaptation material and test material were collected in two different sessions. Results show that a digit accuracy close to 98% can be achieved when the talker is at 1.5 m distance from the array. This result has to be compared with 99.5% accuracy obtained by using a close-talk microphone.

1. INTRODUCTION

Hands-free continuous speech recognition represents a challenging scenario. In the last years, many experimental activities were devoted to investigate the use of microphone arrays for hands-free continuous speech recognition.

This work concerns the use of a Continuous Density HMM (CDHMM) based speech recognizer trained with a large speech corpus of clean speech material and then adapted with some material similar to that of experimental conditions.

Starting from the signals acquired by means of a linear microphone array system, a Time Delay Compensation (TDC) module provides a beamformed input to the recognizer. The advantage of using a microphone array with respect to a single microphone has been addressed in our previous works [1, 2], where hands-free recognition experiments were carried out in various noisy environments. By performing experiments both on real environment data and on simulated data, those works addressed various aspects such as: variabilities due to talker's position, microphone array configuration, noise and reverberation conditions. Another important result was that phone HMM adaptation based on Maximum A Posteriori (MAP) estimation represents an effective way to further reduce the general mismatch (between training conditions and testing conditions) remaining even after the application of microphone array processing.

The purpose of this work is to extend the latter investigation, by considering another HMM adaptation technique based on Maximum Likelihood Linear Regression (MLLR) and a new experimental task of connected digit recognition, more related to a possible immediate application of this technology.

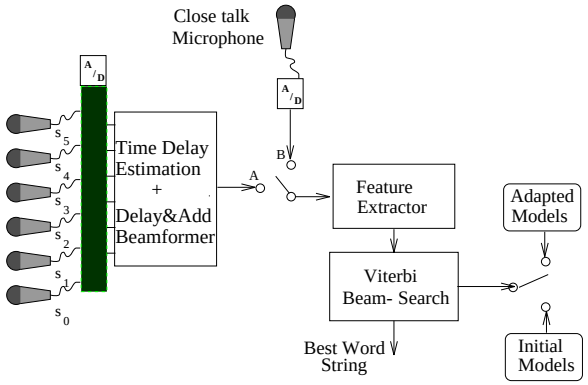

Figure 1: *Block diagram of the recognition system. Two input modalities are included: switch on A corresponds to the use of a six microphone array input, switch on B to close-talk input.*

2. SYSTEM DESCRIPTION

A block diagram of the system is shown in Figure 1. The hands-free recognizer (switch on A) consists of: a linear microphone array module that provides a beamformed output signal; a Feature Extraction (FE) 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 the other way of providing the input signal to the recognizer, that is by using directly a close-talk microphone (switch on B).

2.1. Linear Microphone Array

The use of a microphone array [4] 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 M microphones. 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 [3]. 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 beamformer) [1, 2].

In the following, the analysis is limited to the use of a linear array of six equispaced omnidirectional microphones, characterized by 15 cm distance between adjacent microphones.

2.2. Feature Extraction

The input to the Feature Extractor (FE) is the signal acquired by the close-talk microphone in the case of the baseline system, and the output of the TDC processing when the microphone array is used. The FE input signal is preemphasized and blocked into frames of 20 ms duration (with 50% frame overlapping). 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 utterance. The resulting MCCs and the normalized log-energy, together with their first and second order time derivatives, are arranged into a single observation vector of 27 components.

2.3. 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.

2.4. HMM Adaptation

In automatic speech recognition, performance can drop even dramatically when there is a mismatch between training and test conditions. In the application under investigation this can be due to: the inter-speaker acoustic variability, the acquisition channel and the environmental noise conditions. To cope with this acoustic mismatch, two techniques (originally developed for speaker adaptation) are here examined, based on MAP estimation and on MLLR.

MAP estimation of HMM parameters has been successfully adopted for speaker adaptation. Here, the mean vectors of the Gaussian mixture components are adapted according to a scheme based on MAP estimation [5], while the rest of the model parameters are left unchanged. The speaker-independent clean HMMs are updated as described in previous works [1].

MLLR technique represents another effective solution for adapting an initial set of Gaussian mixture HMMs to a new speaker using a small amount of speaker-specific speech data [7]. MLLR technique was proven to be effective also for compensating acoustic mismatch due to the use of different microphones and environmental acoustic conditions [6].

In this work, MLLR technique is employed for adapting both the means and the variances of Gaussian densities of clean HMMs. The Gaussian densities of the system are grouped into N classes, that are called *regression classes*. Gaussian densities of a regression class are assumed to adapt similarly and to undergo the same transformation. Adaptation of means and variances is performed in two separate steps of an iterative scheme: firstly, the means are adapted; secondly, given the updated means, new variances are computed [6].

For mean adaptation, a linear regression transformation matrix as well as a bias vector are estimated for each regression class, in order to maximize the likelihood of the adaptation data [7]. Note that a full matrix is assumed for all of the regression classes. Gaussian densities associated to a regression class are then adapted by applying the associated transformation matrix and bias vector to the Gaussian mean vectors.

Variance adaptation is performed by estimating a diagonal transformation matrix for each regression class as described in [6] and by using the estimated transformation matrix for updating the diagonal covariance matrices of the Gaussian densities.

Since a set of fixed regression classes has been adopted, when a small amount of adaptation material is available, a robust transformation may not be determined for all the classes. To deal with this specific situation, a global transformation, associated to a regression class formed by all the Gaussian densities of the system, is estimated and used for the classes characterized by lacking of data.

Note that in the experiments described below, a single iteration of the mean and variance adaptation scheme was performed; furthermore, two sets of regression classes were considered (N=1 and N=8).

3. MULTICHANNEL SPEECH CORPUS

Speech material was collected in an office of ($5.5m \times 3.6m \times 3.5m$) size, 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 acoustic source positions.

During a first recording session, 20 phonetically rich sentences (this set does not include any digit; the total number of word occurrences was 210), as well as 30 connected digit strings (consisting in a total of 120 digit occurrences) were uttered by each of four speakers (2 males and 2 females) in a frontal position (F150) at 1.5 m distance from the array. After two days, a new recording session was conducted in the same office, under similar environmental noise conditions: in this case, each of the four speakers uttered 50 connected digit strings (400 digit occurrences), both in the same frontal position (F150) and in a lateral one at 2.5 m distance (L250).

Multichannel recording of each utterance was accomplished by using both a close-talk cardioid microphone and the linear microphone array. Distance between the talker's mouth and the CT microphone was approximately $15cm$.

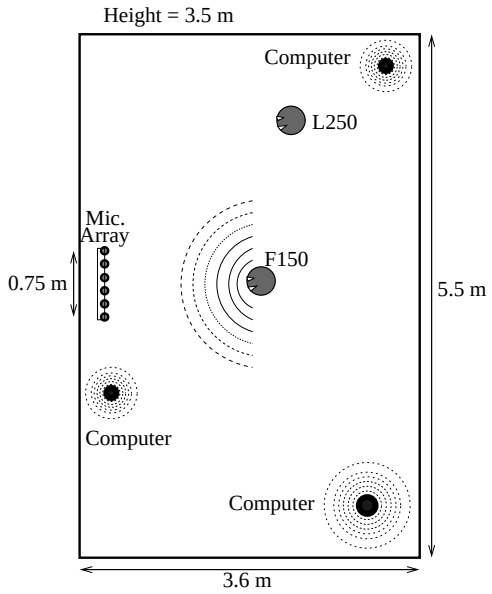

Figure 2: *Map of the experimental room (5.5m x 3.6m x 3.5m), showing the positions of talker (F150, L250), microphone array and computers.*

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 in the range between 12 dB and 18 dB in the case of frontal acquisition (F150) and in the range between 9 dB and 15 dB in the case of lateral acquisition (L250). 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, two development sets and a test set were defined, that consisted in the 20 phonetically rich sentences, 30 digit strings, and 50 digit strings, respectively. Each development set was used to adapt the speaker-independent phone HMMs to acquisition channel, environmental condition and speaker. Performance given in the following is represented as Word Recognition Rate (WRR %) measured on the test set consisting of the 200 strings uttered by the four speakers (resulting in 1600 digit occurrences).

Figure 3 shows performance obtained when the talker was in the frontal position (F150), using the array (*Arr*) or the first microphone (*Mic1*) as input to the recognizer front-end. In both cases, three adaptation techniques were experimented, namely: MAP, $MLLR1m$ (MLLR mean adaptation with 1 class), $MLLR8m$ (MLLR mean adaptation with 8 classes). For every technique, five different sets of adaptation material were adopted, with size ranging between 1 and 20 phonetically rich sentences.

As a reference result, *CT-no ad.* indicates a 99.5 % WRR performance obtained using the close-talk microphone input, without any adaptation; this may represent an upper bound for the other experiments, described in the following.

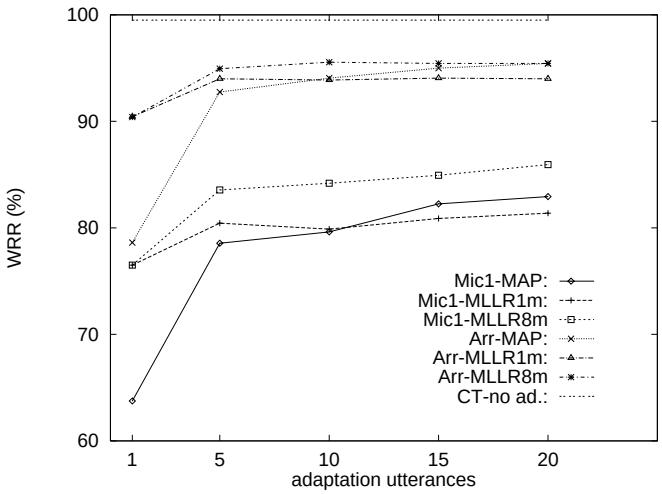

Figure 3: *Experimental results using the close-talk microphone CT, a single far microphone Mic1, or the array Arr as input to the recognizer, given the talker position F150. Results are given for different adaptation techniques and adaptation material size.*

Note that system performance, without any adaptation, using either the microphone array module or the single microphone as input to the recognizer, were 74.3% WRR and 52.5% WRR, respectively.

$N.of Adapt. Utter.$	*1*	*5*	*10*	*15*	*20*
MAP	78.6	92.8	94.0	95.0	95.4
$MLLR1m$	90.4	94.0	93.9	94.1	94.0
$MLLR1mv$	90.5	94.6	94.1	94.2	94.4
$MLLR8m$	90.4	94.9	95.6	95.4	95.4
$MLLR8mv$	90.5	94.9	96.1	96.2	96.2

Table 1: *Performance of the array-based recognizer, for different adaptation techniques and adaptation material sizes, given the talker position F150.*

In the same Figure (as well as in Table 1), one can note that MAP and $MLLR8m$ provides same performance (95.4 % WRR) when using the array as input, while $MLLR8m$ outperforms the other ones, in the case of one microphone input. Actually, this result depends on the adaptation material size. For small adaptation material sets, the MAP technique does not ensure performance comparable to MLLR based ones; this is due to the fact that the basic MAP adaptation scheme, here adopted, does not update model parameters for which no adaptation data are observed.

The advantage of using 8 classes in the MLLR adaptation ($MLLR8m$) with respect to the use of a single global class is also evident. As a final remark, using a very small amount of adaptation material (e.g. one utterance) the adoption of MLLR with 8 classes reduces to the case of using a single global class (see end of Section 2).

A second set of experiments was conducted using the

MLLR adaptation based on varying both means and variances ($MLLR1mv$ and $MLLR8mv$). In this case (see Table 1), 96.2% WRR was obtained using all the adaptation material sentences. Using 10 or more adaptation sentences, $MLLRmv$ shows a marginal benefit with respect to $MLLRm$.

N.of Adapt.Utter.	*1*	*5*	*10*	*15*	*20*
MAP	69.6	89.9	91.7	91.7	91.9
MLLR1m	84.6	90.1	90.3	89.9	90.3
MLLR1mv	86.7	90.2	90.8	90.9	90.2
MLLR8m	84.6	91.2	92.6	92.2	92.6
MLLR8mv	86.7	93.2	93.7	93.6	93.4

Table 2: *Performance of the array-based recognizer, for different adaptation techniques and adaptation material sizes, given the talker position L250.*

Table 2 shows that an analogous trend was obtained for the talker located in position L250. As a reference result, in this case system performance, without any adaptation, using the microphone array module or the single microphone as input to the recognizer, were 57.2% and 40.4%, respectively. It is worth noting that adaptation for this talker position was performed using material collected when the talker was in the other position (F150). As a preliminary result, in this case a drop from 96.2% to 93.4% WRR was found using the $MLLR8mv$ adaptation technique. In other words, this experiment addressed at the same time the misalignment due both to different recording sessions and to different talker positions. The reason for doing this experiment was that previous studies [1] had showed a major influence (on system performance) of talker-array distance than of talker-array direction. Next activities will be devoted to better investigate the use of the same adaptation material for different positions and different recording sessions. We consider these aspects very important for a generalization of the results presented here and, as a consequence, for flexibility of any resulting hands-free recognizer.

		*Mic*1		*Array*	
		*F*150	*L*250	*F*150	*L*250
Phon. rich adapt.	*MAP*	82.9	80.6	95.4	91.9
	MLLR8m	85.9	83.6	95.4	92.6
	MLLR8mv	89.4	88.5	96.2	93.4
Conn. digit adapt.	*MAP*	88.9	86.5	97.3	95.6
	MLLR8m	89.3	84.3	96.7	92.6
	MLLR8mv	93.6	90.8	97.8	95.6

Table 3: *Experimental results for different adaptation techniques and adaptation material sets.*

Finally, an investigation was conducted to focus on the dependency of the adapted system from the adaptation material characteristics. The adaptation material set consisting of digit strings, was used instead of phonetically rich sentences. As it could be expected, results (given in Table 3) show that there is a general noticeable advantage in using a task-dependent adaptation material. Using the array as input, MAP and MLLR adaptation provided 97.3% and 97.8% WRRs, respectively.

5. CONCLUSIONS AND FUTURE WORK

This paper has addressed the generic problem of hands-free continuous speech recognition with a small vocabulary. Many scenarios may be envisaged where this technology could be applied among which: voice commands (automotive, television control, banking, elderly and handicapped assistance, manufacturing process control, surgery, etc.), dictation/data entry for document creation, surveillance.

Several issues remain to be addressed to extend the use of these techniques to more complex situations. In the next future, a particular attention will be devoted to adaptation techniques that can be applied, while the system is on-line, in an unsupervised manner.

Another activity will consist in considering the new experimental task of hands-free dictation of journal articles (Sole24Ore), under development at IRST labs.

6. REFERENCES

[1] 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.

[2] D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer, "Experiments of Speech Recognition in a Noisy and Reverberant Environment using a Microphone Array and HMM Adaptation", *Proc. of EUROSPEECH*, Rhodes, September 1997, pp. 347–350.

[3] 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.

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

[5] 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.

[6] M. J. F. Gales, P. C. Woodland, "Mean and variance adaptation within the MLLR framework", Computer Speech and Language, Vol. 10, pp. 249–264, 1996.

[7] C. J. Leggetter and P. C. Woodland, "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression", *Proc. ICSLP*, Yokohama, September 1994, Vol. 1, pp. 451–454.