TRAINING OF HMM WITH FILTERED SPEECH MATERIAL FOR HANDS-FREE RECOGNITION

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

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

ABSTRACT

This paper addresses the problem of hands-free speech recognition in a noisy office environment. An array of six omnidirectional microphones and a corresponding time delay compensation module are used to provide a beamformed signal as input to a HMM-based recognizer.

Training of HMMs is performed either using a clean speech database or using a filtered version of the same database. Filtering consists in a convolution with the acoustic impulse response between speaker and microphone, to reproduce the reverberation effect. Background noise is summed to provide the desired SNR. The paper shows that the new models trained on these data perform better than the baseline ones.

Furthermore, the paper investigates on MLLR adaptation of the new models. It is shown that a further performance improvement is obtained, allowing to reach a 98.7% WRR in a connected digit recognition task, when the talker is at 1.5 m distance from the array.

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 this purpose.

This work concerns the use of a Hidden Markov Model (HMM) based speech recognizer trained with a corpus of speech material obtained from clean speech signals preprocessed in order to reproduce realistic reverberation and noise effects.

Starting from the signals acquired by means of a linear microphone array, a time delay compensation 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 [1], 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 either on Maximum A Posteriori (MAP) estimation or on Maximum Likelihood Linear Regression (MLLR) represents an effective way to reduce the residual mismatch (between training conditions and testing conditions) persisting after the application of microphone array processing.

In [3], a connected digit recognition task was addressed. Results showed the convenience of using MLLR, especially when only a small adaptation material set is available, and the robustness of the resulting hands-free recognition system when the talker position changes. A limitation of such an approach is due to the fact that the available set of HMMs is trained on clean speech material. Starting from a system tuned on a completely different environmental context, we would like to "move" HMMs toward the real environment context with a few sentences as reference.

The approach followed in this paper consists in trying to exploit some information on the real environment such as impulse response of the room and background noise level to generate a filtered version of the clean speech corpus available for system training. The obtained filtered speech corpus should better match the acoustic operating conditions and allow training of more robust HMMs. Furthermore, the set of models so obtained represents a more suitable initial set for a subsequent model adaptation phase.

This work describes the impulse response measurement procedure as well as the method of filtering speech material, necessary to re-condition clean speech material for HMM training. Experimental results of connected digit recognition are related to the same task described in [3], of which this paper represents an extension.

2. SYSTEM DESCRIPTION

The hands-free recognition system considered in this work consists of: a linear microphone array module that provides a beamformed signal; a Feature Extraction (FE) module; a HMM-based recognizer that can operate either with clean or adapted models (a block diagram of the system can be found in [3]). The same recognition engine was adopted when using a close-talk microphone as input to the FE module.

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. This is accomplished by means of a beamforming technique based on Crosspower Spectrum Phase (CSP) [5] that performs Time Delay Compensation (TDC) [1, 2].

In the following, the analysis is limited to the use of a linear array of six equispaced (at 15 cm) omnidirectional microphones.

2.2. Recognition System

The input to the feature extractor 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.

The HMM module 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.

3. HMM TRAINING

HMM training was accomplished through the standard Baum-Welch training procedure. For the baseline system, HMM training was carried out exploiting a phonetically rich italian corpus, called APASCI [6], acquired in a quiet room (SNR $\geq$ 40 dB) by means of a high quality close-talk microphone. Training set consists of 2166 utterances collected from 100 speakers (50 males and 50 females).

3.1. Training using filtered clean speech

A database of acoustically realistic multichannel signals was artificially recreated, starting from the corpus of single-channel clean signals available for training and from some information on the real operating environment, such as the room impulse response and the background noise level. The obtained filtered version of the training corpus was then used for HMM training.

3.1.1. Impulse response measurement

The time-stretched pulse proposed by Aoshima [7] and generalized in [8] is a chirp-like signal having a flat overall power spectrum, that enables a very accurate measurement of the acoustic impulse response. As a consequence of its extended time duration, this excitation can deliver a large amount of energy, while avoiding problems of dynamic range. The pulse is defined on the discrete frequency domain as the N-point sequence:

$$P(k) = \begin{cases} exp(j2m\pi k^2/N^2) & 0 \leq k \leq N/2 \\ P^*(N-k) & N/2 \leq k \leq N \end{cases} \qquad (1)$$

The parameter m is an integer that determines the stretch of the pulse. The inverse DFT of $P(k)$ is a chirp-like sequence $p(n)$ (see Figure 1) that can be transduced by a loudspeaker into an acoustic signal.

A noteworthy characteristic of $p(n)$ is that its autocorrelation is an almost perfect Dirac delta function. As a consequence, the sequence $y(n)$ acquired by a microphone when the loudspeaker generates this excitation can be easily deconvolved by simply cross-correlating it with the original sequence $p(n)$. The result is the acoustic impulse response from the loudspeaker to the microphone. Apart from the contribution of the frequency response of the loudspeaker, this is exactly the impulse response of the acoustic channel in the acquisition of a talker speaking at the same location of the loudspeaker.

This method was employed to determine the impulse responses from each talker position to each of the six microphones of the array (Figure 2). The overall impulse response of the array-beamformer was then derived by properly delaying and summing the individual impulse responses, according to the aiming of the beamformer toward the talker.

3.1.2. Filtered speech generation

Starting from the clean signals, the reverberation effect of the room on the various channels was reproduced by convolution with the measured impulse responses. The effect of background noise was accounted for by summing the noise recorded inside the room, with properly scaled amplitude to reproduce the desired Signal to Noise Ratio (SNR).

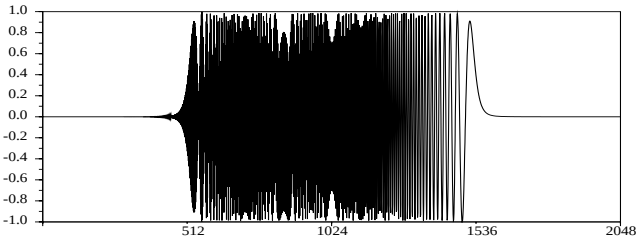

Figure 1: *Example of time-stretched pulse p(n), obtained as inverse DFT of the sequence P(k) in equation (1). Here N=2048, m=1024 (the sequence has been circularly shifted to center the pulse).*

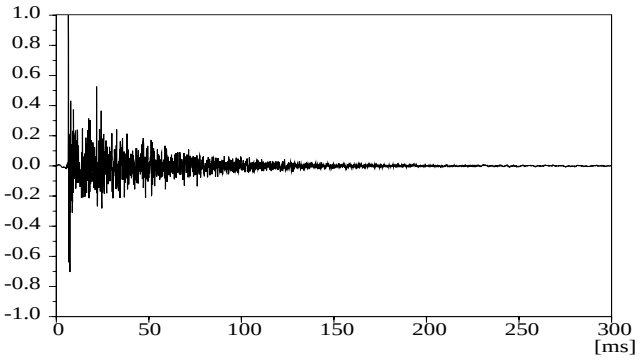

Figure 2: *Example of loudspeaker-to-microphone impulse response measured in the experimental room.*

3.2. HMM Adaptation

In this work, the Maximum Likelihood Linear Regression (MLLR) approach is adopted for adapting an initial set of Gaussian mixture HMMs to new operating acoustic conditions [9, 10].

The Gaussian densities of the system are grouped into 8 regression classes. Adaptation of means and variances is performed, exploiting the adaptation data, in two separate steps of an iterative scheme [10]. Just an iteration of the mean and variance adaptation scheme was performed in the experiments reported in the following [3].

Furthermore, 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. This approach was followed for both mean and variance adaptation [3].

4. MULTICHANNEL SPEECH CORPUS

Speech material was collected (see [3]) in an office of size ($5.5m \times 3.6m \times 3.5m$), characterized by a moderate amount of reverberation ($T_{60} \simeq 0.3s$) as well as by the presence of coherent noise due to some secondary sources (e.g. computers, air conditioning, etc).

During a first recording session, 30 connected digit strings (consisting in a total of 120 digit occurrences) were uttered by each of eight speakers (4 males and 4 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 eight speakers uttered 50 connected digit strings

(400 digit occurrences), both at $F150$ and at the lateral position $L250$ (2.5 m distance, 45 degrees left of the array).

Multichannel recording of each utterance was accomplished by using both a Close-Talk (CT) directional microphone and a linear array of six equispaced (at 15 cm) omnidirectional microphones. Distance between the talker's mouth and the CT microphone was approximately $15cm$. Acquisitions were carried out synchronously for all the input channels at 16kHz sampling frequency, with 16 bit accuracy.

SNR, measured as ratio between average 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 CT microphone signals was in the range between 24 and 33 dB.

5. EXPERIMENTS AND RESULTS

For each speaker, a development set and a test set were defined, that consisted in the 30 connected digit string set and in the 50 connected digit string set, respectively. Each development set was used to adapt speaker-independent phone HMMs to the acquisition channel, to the environmental condition as well as to the speaker.

Performance given in the following is represented as Word Recognition Rate (WRR %), averaged on the test sets of the eight speakers. As a result, each test experiment concerns recognition of 3200 (=400x8) digits.

	$F150$		$L250$	
	Arr	$Mic1$	Arr	$Mic1$
$Baseline$	73.0	50.3	56.6	37.8
$Baseline\_MLLR$	97.8	91.5	95.3	87.8

Table 1: *Recognition performance for the baseline system with and without HMM adaptation.*

Table 1 shows performance (at the two given talker positions $F150$ and $L250$) of the baseline system ($Baseline$) trained using clean speech material. Performance after MLLR adaptation ($Baseline\_MLLR$) is also reported. In particular, considering the talker position $F150$, system performance without any HMM adaptation was 73.0% WRR and 50.3% WRR, using the microphone array module (Arr) or a single microphone of the array ($Mic1$) as input to the recognizer, respectively. As a reference, using CT input, baseline system performance was 99.1% WRR and 99.7%, before and after MLLR adaptation, respectively.

It is worth noting that adaptation experiments for position $L250$ were conducted using material collected at position $F150$.

5.1. HMM training based on filtered speech

The first set of new experiments focuses on the use of HMM models derived from a training on speech material filtered as described in Section 3.

Filtered material can be derived either only adding background noise or also taking into account room acoustics (i.e. by using a measured impulse response). The behavior of system performance depends on the level of noise that is added to clean speech. For this reason a training set was derived adding background noise to the training clean material in order to have signals with a desired SNR. The obtained signals were then used to train a set of HMMs. This operation was repeated for six desired levels of noise in the range 0-25 dB. Each set of models was then used in the recognition tasks: performance is reported in Figure 3. As a result, Figure 3 shows that a maximum of WRR is obtained when additive background

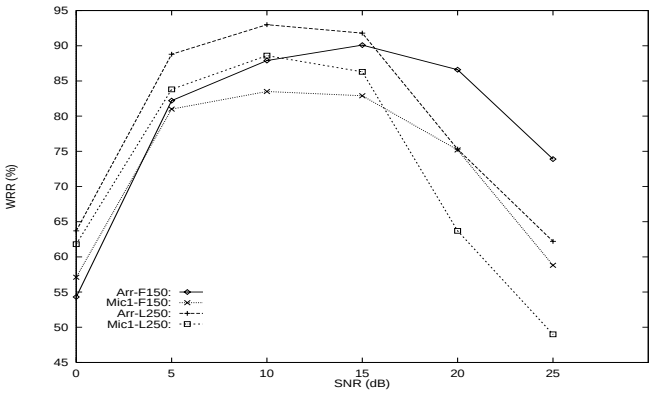

Figure 3: *Recognition performance obtained with different model sets corresponding to training sets characterized by different SNRs.*

noise of 10-15 dB is used. This noise level, as expected, corresponds to that measured in the speech material used for test (see Section 4).

Adding different background noise levels (in this case SNR ranged between 2 and 20 dB) to the clean speech allowed to train a set of HMMs providing performance (84.7% WRR in the case of single remote microphone and 90.3% WRR in the case of microphone array for position $F150$) close to the best ones given in the Figure 3. This training condition, denoted with $Ns2-20dB$, represents an effective and flexible way to derive a robust set of HMMs.

In addition to baseline performance, Table 2 reports on results obtained by the system trained with material derived from clean speech by adding either background noise at 15 dB SNR ($Ns15dB$) or background noise of different levels ($Ns2-20dB$).

	$F150$		$L250$	
	Arr	$Mic1$	Arr	$Mic1$
$Baseline$	73.0	50.3	56.6	37.8
$Ns15dB$	90.1	82.9	91.8	86.3
$Ns2-20dB$	90.3	84.7	90.5	84.7
Ir_Ns15dB	95.6	88.6	94.0	88.6
$Ir\_Ns2-20dB$	95.8	90.6	93.5	89.8
$IrMix\_Ns2-20dB$	94.9	90.1	94.6	89.0

Table 2: *Recognition performance obtained filtering clean speech with different levels of background noise and, in the last three cases, using different room impulse responses.*

When the room impulse response is also used to filter clean speech, a definitely better performance is obtained as shown in Table 2. This advantage is clear both in the case of joint use of impulse response and additive noise at 15 dB SNR (training condition denoted by Ir_Ns15dB) and in the case of joint use of impulse response and additive noise at different SNRs (training condition denoted by $Ir\_Ns2-20dB$). For instance, in the latter case 95.8% WRR is obtained at $F150$ using the microphone array as input. Note that, for all these experiments, impulse response was measured at talker position $F150$.

Another experiment was conducted, where training clean material was filtered adding different background noise levels and using impulse responses measured in four positions (position $F150$ and $L250$ plus two others). Each clean speech signal was filtered choosing randomly a given background noise level between 2 and

20 dB and an impulse response measured in one of the four positions in the room (training condition denoted by $IrMix\_Ns2-20dB$). Recognition performance reported in Table 2 shows that knowing exactly the talker position for an accurate room impulse response helps but it does not seem to be crucial. However, this point requires further investigation.

5.2. HMM adaptation starting from different models

From the previous section, one can observe that any set of models trained with filtered speech material provided results close to baseline with MLLR adaptation.

Another interesting investigation concerns adaptation of models trained with filtered speech. As reported in Table 3, one can observe a further significant improvement with respect to adaptation of the $Baseline$ system. Starting from HMMs trained with the $Ir\_Ns2-20dB$ filtered material 98.7% WRR was obtained at $F150$ using the array as input of the recognizer. It is also interesting to see that even in the case of remote microphone input $Mic1$, the use of filtered speech for training, together with that of MLLR adaptation, represent a convenient approach to improve performance of any hands-free system.

	$F150$		$L250$	
	Arr	$Mic1$	Arr	$Mic1$
$Baseline\_MLLR$	97.8	91.5	95.3	87.8
Ir_Ns15dB_MLLR	98.3	93.8	96.3	92.2
$Ir\_Ns2-20dB\_MLLR$	98.7	95.7	97.9	94.9

Table 3: *Recognition performance with model adaptation starting from HMMs trained with clean or filtered material.*

Experiments discussed above regard the use of 30 connected digit sentences for adaptation. As a final investigation, it was analyzed how this performance changes when a smaller amount of adaptation material is available. Figure 4 reports on some recognition results for position $F150$.

The figure shows that adapting models, trained with $Ir\_Ns2-20dB$ filtered material (curves $Ir\_Ns2-20dB\_Mic1$ and $Ir\_Ns2-20dB\_Arr$), always leads to better results than adapting baseline models (curves $Baseline\_Mic1$ and $Baseline\_Arr$). Furthermore, results show that model adaptation is effective even with just 8 adaptation utterances.

6. CONCLUSIONS AND FUTURE WORK

In hands-free speech recognition the application of microphone array processing compensates only for part of the mismatch between training and testing acoustic conditions.

Given a HMM-based speech recognizer, adaptation of the model set to the new acoustic conditions can be accomplished by exploiting a small amount of adaptation data collected in the operating environment. In this paper a different approach has been presented, and tested in the context of hands-free continuous speech recognition with a small vocabulary. A filtered version of a clean speech corpus is derived and used for HMM training. The resulting models are therefore conditioned to the operating acoustic conditions that have been assumed. Recognition experiments showed that these new models ensure performance clearly better than that obtained with baseline models and close to that obtained adapting baseline models. Furthermore, training models with filtered speech results in a good initial set of models for a subsequent adaptation phase.

In the future the use of models trained with filtered speech will be investigated in combination with on-line adaptation.

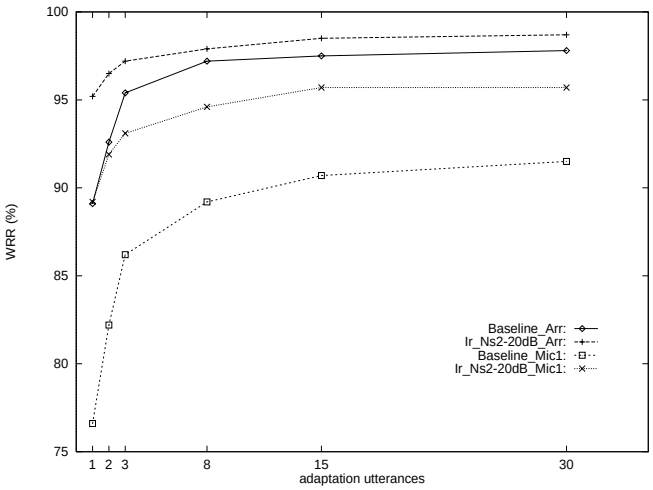

Figure 4: *Adaptation results with different amount of speech data for speakers in position $F150$, using a single microphone or the microphone array as input. Initial models are either baseline models (curves denoted with $Baseline\_Mic1$ and $Baseline\_Arr$) or models trained under $Ir\_Ns2-20dB$ condition (curves $Ir\_Ns2-20dB\_Mic1$ and $Ir\_Ns2-20dB\_Arr$).*

7. REFERENCES

[1] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, "Microphone Array based Speech Recognition with different talker-array positions", *Proc. of ICASSP*, 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*, September 1997, pp. 347–350.

[3] D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer, "Experiments of HMM Adaptation for Hands-Free Connected Digit Recognition", *Proc. of ICASSP*, May 1998, pp.I-473-476.

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

[5] 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.

[6] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus", In *Proc. of ICSLP*, September 1994, pp. 1391–1394.

[7] N. Aoshima, "Computer-generated pulse signal applied for sound measurement", J. Acoust. Soc. Am., vol. 69(5), pp. 1484-1488, May 1981.

[8] Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses", J. Acoust. Soc. Am., vol. 97(2), pp. 1119-1123, February 1995.

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

[10] 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.