AN ON-LINE ACOUSTIC COMPENSATION TECHNIQUE FOR ROBUST SPEECH RECOGNITION

Diego Giuliani

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
38050 Povo, Trento, Italy
giuliani@itc.it

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

10.21437/Eurospeech.1999-543

ABSTRACT

In this work we report on the use of an on-line acoustic compensation technique for robust speech recognition. With this technique acoustic mismatch between training and actual conditions is reduced through acoustic mapping. At recognition stage, observation vectors delivered by the acoustic front-end are mapped into a reference acoustic space, while input data are exploited to update the statistical parameters of the mapping. Experimental results, obtained for matched and unmatched training and testing environment conditions, show that the investigated technique tangibly improves the performance of a speaker independent speech recognizer based on hidden Markov models. Furthermore, recognition results are close to those obtained with unsupervised incremental model adaptation based on maximum likelihood linear regression.

Keywords: acoustic mapping, acoustic compensation, robust speech recognition.

1. INTRODUCTION

In automatic speech recognition there is often a gap between the performance obtained in laboratory and that obtained in operating conditions. One major reason of this performance gap may lie in the acoustic mismatch between training and operating conditions. Acoustic mismatch may be due to different causes such as intra- or inter-speaker acoustic variability, characteristic of microphone and transmission channel, noise and reverberation conditions, the position of the microphone with respect to the speaker in hands-free speech recognition. Different approaches have been developed in the last years to tackle this problem [6, 4, 5, 12, 11, 8]. Some of them deal with a challenging scenario in which, at recognition stage, the recognition system is continuously adapted to the current acoustic conditions exploiting past input data [9, 10]. In this work we deal with such an *on-line* adaptation scenario.

A distinction can be done among adaptation approaches which modify the parameters of the speech recognizer, map input observation vectors or both [14, 15]. In this work we investigate the use of an on-line acoustic compensation technique based on acoustic mapping for compensating acoustic mismatch between training and testing conditions. At recognition stage, observation vectors delivered by the acoustic front-end are mapped into a reference acoustic space, while statistics for performing acoustic mapping are collected exploiting input data. This technique, firstly introduced in [7], is independent of the particular speech recognizer used.

The investigated technique can be thought as an on-line version of the "blind RATZ" algorithm introduced by Moreno et al. [11] for compensating the acoustic mismatch between clean and noisy environments. In that approach the clean speech statistics are characterized by means of a Gaussian mixture density and it is assumed that the effect of the environment can be modeled in terms of variations induced on the parameters of this Gaussian mixture density. In practice, parameters of a Gaussian mixture density are first estimated using clean speech and then re-estimated using noisy data. Variations induced by noisy speech on means and variances of this reference Gaussian mixture are then used, during recognition, for performing acoustic compensation. An additive compensation term is computed for each input noisy observation vector before recognition with a speech recognizer trained on clean speech. We implement the above approach in such a way that, at recognition stage, not only compensation is performed but also variations induced on the parameters of the reference Gaussian mixture density are estimated. For this purpose, at recognition stage, the reference Gaussian mixture density, whose parameters are estimated using clean speech, is incrementally adapted, through maximum likelihood linear regression (MLLR) [10, 4], to the current environment conditions and to the current speaker.

Results of speech recognition experiments show that the proposed technique tangibly improves the performance of a speaker independent speech recognizer based on hidden Markov models (HMMs) when training and testing environment conditions either match or do not match. Furthermore, recognition results are close to those obtained with unsupervised incremental HMM adaptation based on MLLR.

2. COMPENSATION APPROACH

Observation vectors delivered by the acoustic front-end are mapped into the reference acoustic space with the aim of reducing acoustic mismatch.

2.1. Gaussian Mixture Model

Let us assume that the data set, available to train the speech recognizer, is drawn from a reference acoustic space represented by a Gaussian mixture probability density function having the following parametric form:

$$p(\mathbf{o}) = \sum_{k=1}^{M} \omega_k \mathcal{N}(\mathbf{o}; \mu_k, \mathbf{\Sigma}_k) \qquad (1)$$

where $\mathbf{o}$ is a d-dimensional acoustic observation vector, M is the number of Gaussian components and $\mathcal{N}(\mathbf{o}; \mu_k, \mathbf{\Sigma}_k)$ denotes a Gaussian density function with mean vector μ_k and covariance matrix $\mathbf{\Sigma}_k$. In addition, $\{\omega_k\}$ are the non-negative mixing parameters which satisfy the constraint $\sum_{k=1}^{M} \omega_k = 1$.

Parameters of this Gaussian mixture can be estimated by means of the EM algorithm [2, 13] exploiting the available training set. Consider another data set, drawn from a new acoustic space that can be considered as a "distorted" or/and a "noisy" version of the reference space. Also for the new space we can assume a Gaussian mixture density as follows:

$$p(\hat{\mathbf{o}}) = \sum_{k=1}^{M} \hat{\omega}_k \mathcal{N}(\hat{\mathbf{o}}; \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k) \tag{2}$$

where $\hat{\mathbf{o}}$ is an observation drawn from the new space. Parameter estimation can be accomplished exploiting speech data from the new acoustic space. An important issue when using the EM algorithm is the initialization of the parameters, since the initialization not only affects convergence of the algorithm but also the final estimates [2, 13]. An interesting case occurs when the estimates for the parameters of the reference Gaussian mixture are used as initial estimates for the parameters of the Gaussian mixture modeling the new acoustic space. Under certain assumptions, variations induced both on initial mean vectors and on covariance matrices can be interpreted as the effect of "distortion" or/and of "noise" in the reference acoustic space [11]. These variations can be used for deriving an acoustic mapping between the new and the reference acoustic spaces [11].

2.2. Acoustic Mapping

Given an observation vector $\hat{\mathbf{o}}$ from the new acoustic space, it is mapped into the reference space by the following vector function:

$$f(\hat{\mathbf{o}}) = \hat{\mathbf{o}} + \sum_{k=1}^{M} \alpha_k(\hat{\mathbf{o}}) \mathbf{\Delta}_k \tag{3}$$

where the weighting coefficient $\alpha_k(\hat{\mathbf{o}})$ is derived by application of the Bayes rule as

$$\alpha_k(\hat{\mathbf{o}}) = \frac{\hat{\omega}_k \mathcal{N}(\hat{\mathbf{o}}; \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k)}{\sum_{i=1}^{M} \hat{\omega}_i \mathcal{N}(\hat{\mathbf{o}}; \hat{\mu}_i, \hat{\mathbf{\Sigma}}_i)} \tag{4}$$

and $\mathbf{\Delta}_k$ is defined as

$$\mathbf{\Delta}_k = \mu_k - \hat{\mu}_k. \tag{5}$$

Note that the additive compensation term in the right hand side of Eq. (3) is obtained as a linear combination of the mean vector differences $\{\mathbf{\Delta}_k\}$ with weighting coefficients depending on $\hat{\mathbf{o}}$.

2.3. On-line Learning

When the acoustic conditions holding at recognition stage are unknown, the approach described above can still be used. At recognition stage, variations induced on the parameters of a reference Gaussian mixture by the input utterances can be used for defining acoustic mapping between the new and the reference spaces. The Gaussian mixture model defined by Eq. (1) can be thought as a single state hidden Markov model having a Gaussian mixture emission density. Theoretical frameworks developed for incremental (or on-line) adaptation of continuous density HMMs can be applied. In this sense, on-line Bayesian learning [9] and incremental MLLR [10, 4] are viable approaches. Therefore, at recognition stage, the reference Gaussian mixture density can be incrementally adapted toward the current acoustic conditions. The variations induced by past input data on its initial parameter values allow for dynamically specifying a suitable acoustic mapping for incoming utterances (see Section 3.2).

3. EXPERIMENTAL SETUP

Two sets of recognition experiments were carried out using an HMM based speech recognizer trained with clean speech. The first set of experiments concerned a dictation task with speech collected by means of a close-talk microphone. The second set of experiments concerned hands-free connected digit recognition. For the dictation task training and testing acoustic conditions matched (i.e. acquisitions with the same microphone type and high signal-to-noise ratio) while for the hands-free connected digit task they were very different. For comparison purposes, experiments concerning incremental unsupervised model adaptation were also carried out. In all the experiments it was assumed that session boundaries through different speakers were known as well as the utterance order.

3.1. Acoustic Models

The acoustic front-end computed, for each frame of speech signal, 12 mel scaled cepstral coefficients (MSCCs) plus the log-energy. These coefficients, together with their first and second order time derivatives, were arranged into a 39-dimensional feature vector. MSCCs were normalized through mean subtraction on a sentence basis. Furthermore all 39 features were scaled in order to obtain unit variances on the training set of the speech recognizer. The speech recognition system, for the Italian language, was based on a set of continuous density HMMs related to 48 SAMPA units, plus a model for silence. Output probability distributions were modeled with mixtures of 32 Gaussian density functions having diagonal covariance matrices. Speaker independent HMMs were obtained using the Italian clean speech corpus APASCI [1] as training material.

3.2. Compensation Module

The parameters of two Gaussian mixtures, both having 128 components with diagonal covariance matrices, were estimated by means of the EM algorithm using clean speech. The first Gaussian mixture was trained using the clean speech, task independent, corpus APASCI. The other one was trained exploiting a clean speech corpus consisting of 900 utterances concerning repetitions of digit sequences by 36 speakers (19 males and 17 females). This second Gaussian mixture allowed to model a task dependent reference acoustic space for connected digit recognition. During training, mixing parameters were maintained fixed and equal among them.

At recognition stage, the reference Gaussian mixture modeling the clean speech statistics was adapted through incremental MLLR [10] in order to model the current acoustic space. For this purpose Gaussian components of the reference mixture were grouped into 4 static regression classes. For each regression class a block diagonal transformation matrix (with blocks corresponding to static features and first and second order time derivatives,

respectively) was assumed for adapting the means, while a diagonal transformation matrix was adopted for adapting the variances [4]. After each input utterance, means and variances were updated allowing to specify acoustic mapping, as described in Section 2.2, for the next incoming utterance. In order to accelerate the adaptation process, transformation matrices associated with a global regression class (formed by all the Gaussian components in the mixture) were estimated and employed for updating parameters of Gaussian components associated with regression classes for which insufficient statistics were collected at a given point. An empirical threshold was adopted to decide when enough data were observed for robust estimation of transformation matrices associated with a given regression class.

3.3. Incremental Model Adaptation

Incremental unsupervised MLLR represents an effective way to progressively adapt an initial set of continuous density HMMs to the current speaker and environment conditions [10, 4]. Incremental unsupervised adaptation relies on recognition results and therefore wrong text transcriptions may affect the adaptation process. At recognition stage, means and variances of Gaussian components in the system were adapted using transformation matrices associated with 8 static regression classes. Block diagonal transformation matrices (with blocks corresponding to static features and first and second order time derivatives, respectively) were assumed for adapting the means, while diagonal transformation matrices were adopted for the variances [4]. Model parameters were updated every 10 seconds of input speech. As described in Section 3.2, transformation matrices associated with a global regression class were estimated and adopted for regression classes for which insufficient statistics were collected at a given point, in order to reduce the number of recognized sentences before any transform matrices may be robustly estimated.

3.4. Dictation Task

This task was a continuous speech task concerning the dictation of fragments of newspaper articles taken from the financial Italian journal "Il Sole 24 Ore" [3]. The task had a closed dictionary consisting of 10,000 words. The recognizer adopted a bigram language model and an efficient beam search strategy [3]. 30 utterances were acquired in office environment, using a close-talk microphone, from each of 12 test speakers. Uttered texts were different speaker by speaker and on average each speaker uttered 600 words.

Utterances	*Baseline*	*Comp*	*Adapt*
1-10	89.0	89.7	89.9
11-20	87.8	90.6	90.7
21-30	91.0	92.7	93.3
1-30	89.2	91.0	91.3

Table 1: *Word recognition rates (%) for the dictation task obtained with the baseline system ("Baseline" column), performing acoustic compensation ("Comp" column) and performing unsupervised incremental adaptation ("Adapt" column). Results are reported for three blocks of utterances in sequence (utterances 1-10, 11-20 and 21-30) and the whole set of data (utterances 1-30).*

Table 1 reports the word recognition rates (WRRs) achieved with the baseline system ("Baseline" column) and performing acoustic compensation before recognition ("Comp" column). For acoustic compensation the reference Gaussian mixture trained with the APASCI corpus was used. Reported WRRs were computed averaging performance over the 12 speakers. In order to better analyze the effect of the acoustic mapping in time, results are presented for three blocks of utterances in sequence (utterances 1-10, 11-20 and 21-30) and for the whole set of data (utterances 1-30). Marginal benefits can be noted even in the first block of utterances (utterances 1-10). For comparison purposes, recognition results obtained performing incremental model adaptation are also reported in Table 1 ("Adapt" column). Being the test material clean speech, results show how both the proposed technique and incremental model adaptation allow the system to progressively adapt to the speaker.

3.5. Hands-free Recognition Task

Hands-free speech recognition experiments concerned connected digit recognition for which a suitable speech corpus was employed [8]. Multichannel recordings were accomplished in an office by using both a close-talk cardioid microphone and a linear microphone array. 50 sentences, concerning eight digit strings, were uttered by 8 speakers (4 males and 4 females) in a frontal position at 1.5 m distance from the microphone array (position F150) and in a lateral position at 2.5 m distance and 45° angle from the array (position L250). Each speaker uttered a total of 400 digits in each position. Experiments were carried out by considering speech signals collected with a single microphone located in the central position of the array.

Utterances	*F150*		
	Baseline	*Comp*	*Comp_td*
1-15	57.2	74.3	78.0
16-30	57.2	80.0	88.0
31-50	52.3	80.9	87.0
1-50	55.3	78.7	84.6

Table 2: *Word recognition rates (%) for the hands-free connected digit task obtained with the baseline system ("Baseline" column) and performing acoustic compensation ("Comp" and "Comp_td" columns). Results are reported for the talker position F150 considering three blocks of utterances in sequence (utterances 1-15, 16-30 and 31-50) and the whole set of data (utterances 1-50).*

Utterances	*L250*		
	Baseline	*Comp*	*Comp_td*
1-15	53.7	70.7	77.8
16-30	54.0	79.7	89.8
31-50	46.7	76.4	86.8
1-50	51.0	75.7	85.0

Table 3: *Word recognition rates (%) for the hands-free connected digit task obtained with the baseline system ("Baseline" column) and performing acoustic compensation ("Comp" and "Comp_td" columns). Results are reported for the talker position L250 considering three blocks of utterances in sequence (utterances 1-15, 16-30 and 31-50) and the whole set of data (utterances 1-50).*

Table 2 reports the WRRs achieved for talker position F150 with the baseline system ("Baseline" column) and performing acoustic compensation with task independent

and task dependent reference Gaussian mixtures ("Comp" and "Comp_td" columns, respectively). WRRs were computed averaging performance over the 8 speakers. Results are presented for three blocks of utterances in sequence (utterances 1-15, 16-30 and 31-50) and for the whole set of data (utterances 1-50). As expected, using the task dependent reference Gaussian mixture ensures better recognition performance. This is confirmed by experiment results reported in Table 3 for position L250. When the task dependent reference Gaussian mixture is used for acoustic compensation, recognition results for position L250 are even better than those obtained for position F150.

Note that the WRRs achieved with the baseline system for clean speech acquired with the close-talk microphone in position F150 and L250 were 98.9% and 98.7%, respectively.

Utterances	*F150*		*L250*	
	Comp_td	*Adapt*	*Comp_td*	*Adapt*
1-15	78.0	78.0	77.8	74.8
16-30	88.0	91.4	89.8	93.3
31-50	87.0	91.9	86.8	91.6
1-50	84.6	87.6	85.0	87.0

Table 4: *Word recognition rates (%) for the hands-free connected digit task obtained performing acoustic compensation ("Comp_td" columns) and incremental unsupervised model adaptation ("Adapt" column). Results are reported for the two talker positions F150 and L250.*

Recognition results reported in Table 4 show how the proposed technique and incremental model adaptation allow the system to progressively adapt to the current environment conditions and to the speaker ("Comp_td" and "Adapt" columns, respectively). As for the dictation task, incremental model adaptation based on MLLR performs better than the proposed technique.

4. CONCLUSION

In this work we investigated the use of an on-line technique for acoustic mismatch compensation. Experimental results show that the proposed technique is able to tangibly improve the performance of an HMMs based speech recognizer both when training and testing acoustic conditions match and do not match.

In comparison with incremental model adaptation based on MLLR the proposed technique seems to be less effective. However, there are a lot of issues concerning the proposed techniques that deserve to be investigated and that could provide a margin of improvement. Some of them concern the choice of the number of components for the Gaussian mixture modeling the reference acoustic space, the algorithm for adapting the reference Gaussian mixture, the set of acoustic features to be transformed (i.e. static and dynamic features or static features alone).

The proposed technique is attractive from a computational point of view and is independent of the particular type of speech recognizer in use. Furthermore it is suitable to be used in combination with incremental model adaptation. Preliminary experiments in which acoustic compensation is carried out simultaneously with incremental model adaptation are encouraging.

5. REFERENCES

[1] 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 *ICSLP*, pages 1391–1394, Yokohama, 1994.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society Ser. B*, 39:1–38, 1977.

[3] M. Federico, M. Cettolo, F. Brugnara, and G. Antoniol. Language modeling for efficient beam-search. *Computer Speech and Language*, 9:353–379, 1995.

[4] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[5] M. J. F. Gales and S. J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9:289–307, 1995.

[6] J.-L. Gauvain and C.-H. Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.

[7] D. Giuliani. An On-line Technique for Speaker and Environment Adaptation. In *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999.

[8] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer. Training of HMM with Filtered Speech Material for Hands-free Recognition. In *Proc. of ICASSP*, pages I–449–452, Phoenix, March 1999.

[9] Q. Huo and C.-H. Lee. On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate. *IEEE Trans. on Speech and Audio Processing*, 5(2):161–172, 1997.

[10] C. J. Leggetter and P. C. Woodland. Flexible Speaker Adaptation for Large Vocabulary Speech Recognition. In *Proc. of EUROSPEECH*, pages 1155–1158, Madrid, Sept. 1995.

[11] P. J. Moreno, B. Raj, E. Gouvêa, and R. M. Stern. Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition. In *Proc. of ICASSP*, pages I–137–140, Detroit, May 1995.

[12] L. Neumeyer and M. Weintraub. Probabilistic Optimum Filtering for Robust Speech Recognition. In *Proc. of ICASSP*, pages I–417–420, Adelaide, April 1994.

[13] R.A. Redner and H.F. Walker. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM review*, 26(2):195–239, April 1984.

[14] A. Sankar and C.-H. Lee. A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, 4(3):190–202, 1996.

[15] Y. Zhao. An Acoustic-Phonetic Based Speaker Adaptation Technique for Improving Speaker Independent Continuous Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, 2(3):380–394, 1994.