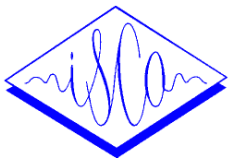

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Channel Adaptation for a Continuous Speech Recognizer *

L. Fissore ◇ P. Laface ⋆ G. Micca ◇ G. Sperto ◇

◇ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy
⋆ Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - 10129 Torino Italy

Abstract

This paper deals with adaptation to new test environments of an HMM recognizer trained in a given condition. In particular, a speaker-independent continuous speech recognizer has been trained using a multispeaker database collected through a telephone with an electret transducer and tested with several types of hand-sets having different spectral response characteristics. The results of these experiments show that system performance drops dramatically on test utterances collected through carbon transducers.

A procedure for hierarchical spectral mapping, originally proposed for speaker adaptation, is used for channel adaptation. A codebook adapted to the new channel is obtained from a database including a limited number of adaptation sentences. Test sentences are then vector quantized using the adapted codebook. Tested with sentences acquired through a carbon transducer, more than 60% of performance improvement is achieved, while nearly 100% of the errors can be recovered when electret type transducers are used for testing.

Adapted HMMs are then estimated by aligning the sentences collected in the new environment against the HMMs trained in the original environment, and nearly 92% of the errors are recovered in the carbon transducer test.

1 Introduction

A speaker-independent recognizer to be used over the PSTN for general user applications should not only be insensitive to speaker variability in terms of sex, age, dialect inflections, speaking rate, but it should also be robust to changes in spectral shaping, to noise and signal levels, to echoes and other channel distortions.

Actually, instead, an automatic speech recognition system tested on a transducer other than the one on which it was trained usually yields dramatic degradations in recognition accuracy [1]. Thus, fast speaker and channel adaptation techniques are being actively investigated to obtain good performances for every speaker, microphone and environment [1, 6]. On-line signal normalization has the advantage that a system can adapt to dynamic variations in the environment.

*This work has been partially supported by ECC Esprit II project 2218 – "SUNDIAL"

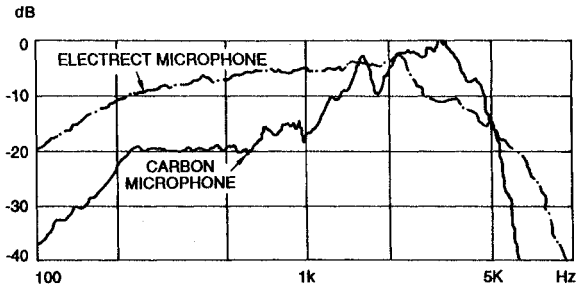

Figure 1: Frequency response of training and test transducers

but simpler and effective off-line procedures can be devised that allow existing large databases to be used for training models related to a new application or simply to adapt them to a new environment.

The paper deals with channel adaptation to new test environments of a speaker-independent, continuous speech recognizer trained in a given condition. Test environments are characterized by different transducers (linear electret or carbon) which provide different spectral shaping of input speech. The training database, instead, has been collected through a single telephone with electret transducer.

Test sentences are vector quantized through an adapted codebook obtained by a codebook mapping procedure using a database with a limited number of adaptation sentences. Adapted HMMs are then estimated by aligning the sentences collected in the new environment against the HMMs trained in the original environment. A linear interpolation is then performed which combines the original and adapted HMMs on a state-by-state basis, according to the number of observations of a given state in the test utterances.

2 Baseline System

The system used in the adaptation experiments is a speaker-independent continuous speech recognizer which is the acoustic Front-End of a Speech Understanding System [4]. Its vocabulary includes 800 words represented in terms of context-dependent units modeled with Discrete Density Hidden Markov Models. Model training is performed by means of the Forward-Backward procedure [2], while recognition is carried out according to a beam search constrained

Initialization (adaptation level $n = 0$)

While $n \leq n_{max}$ do

1. The centroids $a_i, (i = 1, \ldots m = 2^n)$ of the adaptation space are obtained by LBG clustering.
2. Codevectors in the original space C_o, $(c_i, i = 1, \ldots m = 2^{n_{max}})$ are clustered into $m = 2^n$ regions (with centroids $o_i, (i = 1, \ldots m)$) according to the nearest neighbor criterion with respect to the centroids in the adaptation space $a_i, (i = 1, \ldots m)$.
3. The deviation vectors $p_i = a_i - o_i, (i = 1, \ldots m)$ are evaluated.
4. To maintain contiguity between adjacent clusters, the weighting factors $w_{ik} = |c_i - o_k|^{-\alpha}, (i, k = 1, \ldots m)$ are evaluated to obtain a codeword dependent deviation:
$$\delta_i = \frac{\sum_{k=1}^{m} w_{ik} p_k}{\sum_{k=1}^{m} w_{ik}} \quad (1)$$
5. Codevectors c_i are mapped into the new adaptation space according to their deviation: $c_i = c_i + \delta_i, (i = 1, \ldots m = 2^{n_{max}})$
6. $n = n + 1$

End of loop

Table 1: Codebook adaptation procedure

Forward decoding algorithm [3].

The codebooks and the models were obtained using a speech database including 8800 utterances from 110 speakers. The database was collected through a standard telephone, with an electret transducer, connected to a local PABX line. The same telephone and a new one, with a carbon transducer, were used to collect 1200 test utterances from a new set of 20 speakers. In Fig. 1, the transfer function of the electret transducer is compared with the transfer function of the carbon transducer used for the first test experiment.

Recognition rate, in terms of word accuracy, amounted to 66.1% for the electret transducer test set, while only 28.8% was obtained with the carbon transducer test set.

3 Adaptation Schemes

3.1 Codebook Mapping

The number of sentences included in the *carbon* test database amounts to about 13% of the utterances in the training database. It is not big enough to reliably estimate a codebook, otherwise a new set of models could be trained for the new environment. Thus, in the first approach, a procedure for hierarchical spectral mapping, based on the LBG algorithm [7], originally proposed for speaker adaptation[5], has been adapted to map the original codebook to the new acoustic space.

The procedure is recalled in Tab. 1, where C_o is the *original codebook* obtained from the 110 training speakers using the electret transducer.

At level $n = 0$, the deviation vector does not depend on index i: $\delta_i = p_1, \forall\ i$, which corresponds to the deviation between the centroid of the training and of the test data, the former being approximated by the centroid of the original codevectors. Mapping is refined in the successive iterations, while parameter α acts as a contiguity controller. At higher levels, in fact, the deviation vectors p_i of two adjacent codewords might map them into separate regions of the adaptation space. Optimal setting of parameter α is obtained experimentally. Full contiguity is assured by setting parameter α to 0, but large distortion is observed for the corresponding adapted codebook. On the other hand, setting $\alpha = \infty$ enhances the contribution of the actual centroid and of its nearest neighbors to the the deviation vector p_i, but contiguity is not assured. This effect is likely to appear even for $\alpha > 1$, therefore, the optimal α is expected to be within the range $[0, 1]$.

3.1.1 Codebook adaptation results

In the first set of experiments recognition performance sensitivity to parameter α has been evaluated. The whole database collected through the carbon transducer was used both for adaptation and for testing, vector quantized by means of three different codebooks, one for the cepstral parameters, one for the differential cepstral parameters, and the last one for the energy and differential energy taken together. The number of bit used for each codebook were 8, 8, and 5 respectively. The maximum adaptation level was set to 8.

The results of this experiment are summarized in Fig. 2 where word accuracy is given as a function of the adaptation level and for $\alpha = 1.0$, 0.75 and 0.50. The effectiveness of the hierarchical clustering approach increases according to adaptation level provided that parameter α is less than 1.0.

In the second set of experiments, the 20 speaker database was split into two subsets: an adaptation subset (15 speakers) and a test subset (5 speakers). These sets were generated according to the leave one out strategy (hold one set of 5 speaker out in our case) for a total of four independent experiments. The results are given in Tab. 2, where *Electret* refers to an experiment with the same 20 speakers collecting the test sentences through the same electret transducer that was used for the training sentences. This result can be considered, therefore, as an upper bound to recognition accuracy.

Fig. 3 plots word accuracy as a function of the

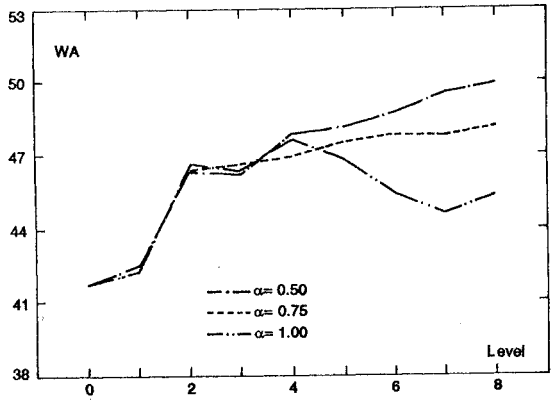

Figure 2: Word accuracy as a function of the adaptation level.

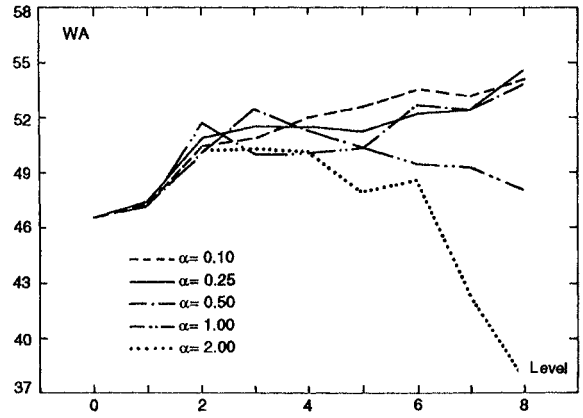

Figure 3: Word accuracy with codebook adaptation to carbon transducer; adapt. sp. 1-15, test sp. 16-20.

adaptation level for different values of parameter α. $\alpha = 0.25$ gave the best performance at the highest adaptation level. A relevant improvement is already observed at level $n = 0$, which corresponds to a constant codevector shift. Further improvement is gained by moving toward higher levels. Nearly 60% of performance degradation with the carbon transducer test (from 66.1 to 28.8) has been recovered by the adaptation scheme.

3.1.2 Results with other channels

The above described procedure was followed in successive experiments, where three other transducers were employed (referred to as **AZ**, **BR** and **FU**) for collecting the test utterances. They are of the same type of the linear-electret transducer (**SI**), used for the training utterances, but with different frequency response characteristics, as shown in Fig. 4.

α = 0.25 level = 8	
Test database	Word accuracy
Carbon	28.8
Carbon with adaptation	51.7
Electret	66.1

Table 2: Codebook adaptation

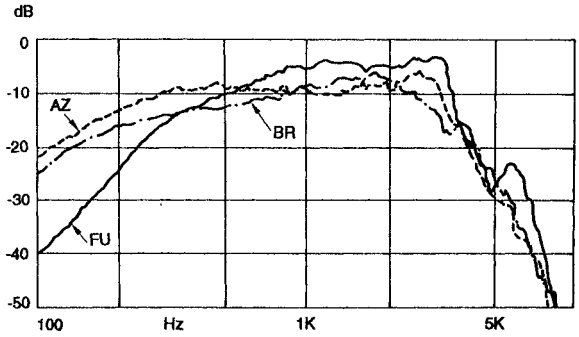

Figure 4: Frequency response of three other test transducers.

Transducer	AZ	FU	BR
Before adaptation	63.8	60.9	63.5
After adaptation	67.3	66.5	67.8

Table 3: Word accuracy with codebook adaptation to electret transducers.

Since the differences among electret-type transducers are less relevant with respect to the carbon transducer, only a slight performance decrease can be observed. Each of 10 new speakers provided 60 sentences. The database was partitioned into an adaptation set including the utterances of 7 speakers, the remaining set of three is used as test set. Average results after codebook adaptation for the three transducers are given in Tab. 3. It is worth noting that, after adaptation, the performance of the worst of the three transducers (**FU**) becomes comparable to that of the other ones.

3.2 Feature adaptation

The codebook adaptation procedure illustrated in Section 3.1 generates a mapping of the original codebook to the test environment. It can be used as well, by swapping the original and the test environments, for adaptation of test features to the original environment. Input frames are, therefore, mapped into the training spectral space and then quantized by means of the original codebook.

Again the 20 speaker database collected through the carbon transducer has been used according to the leave one out strategy. The results of this approach, reported in Tab. 4, show that feature adaptation yields an improvement of 3 percentual points with respect to codebook adaptation. This improvement was not considered statistically significant, and the first approach has been pursued together with a procedure for adaptation of the HMMs.

$\alpha = 0.25$ $\quad$ *level* $= 8$		
adaptation		
Codebook	Features	HMM
51.7	54.7	62.94

Table 4: Word accuracy with codebook, feature and HMM adaptation

3.3 HMM adaptation

Codebook and feature adaptation allows two different channels to be normalized by codevector dependent spectral mapping. The experiments so far illustrated, however, have been performed using the HMMs trained with the original database that was used for codebook generation. Therefore, the acoustic-phonetic information modeled by the HMMs was not exploited to make the models more robust to the new environment.

Since the size of the adaptation database is not large enough to train robust HMMs, linear interpolation [2] can be used to trade-off robust original models M_o and "adaptation" models M_a.

$$M = (1 - \lambda)M_o + \lambda M_a \qquad (2)$$

The "adaptation" models are obtained according to the following procedure:

- Test utterances are vector quantized using the codebook generated through the codebook adaptation procedure.
- State-by-state segmentation is performed through Viterbi alignment of the adapted sequences against the *original* HMMs.
- M_a parameters are estimated by state-dependent occurrence counts.

The weight factor λ is obtained by a simple linear function of the state occurrence counts, with upper and lower thresholds N_1 and N_2 respectively, obtained by partitioning the state counts into three groups: the first group included the lower 30% part of occurrence counts, the second one included the higher 10%, and the third one the intermediate occurrence counts. N_1 was set to the maximum count of the first group, N_2 to the minimum count of the second group. $N_1 = 50$ and $N_2 = 1050$ were eventually obtained. As a consequence, the total set of stated was composed of 280 states from original models, 20 states from adapted ones, while the remaining 604 states were interpolated.

Tab. 4 summarize the results for the three approaches. The performance gap from homogeneous testing conditions and different channel conditions (66.1% and 28.8% in Tab. 2) is almost completely recovered.

4 Conclusions

Three procedures for off-line channel adaptation of a speaker-independent continuous speech recognizer have been evaluated. These methods are effective when the adaptation database is large enough to allow codebook generation and HMM estimation, even though its size is a small fraction of the "original" database. Codebook adaptation, followed by HMM adaptation, recovers nearly than 92% of the errors.

Feature adaptation, on the other hand, provided a small improvement with respect to codebook adaptation and seems best suited to on-line adaptation schemes. In fact, its first adaptation level ($n = 0$) is the basic procedure for spectral mean compensation that can be effectively employed whenever only few seconds of the unknown channel data are available.

References

[1] A. Acero, and R.M. Stern. "Environmental Robustness in Automatic Speech Recognition." ICASSP, pp. 849-852, 1990.

[2] L.R. Bahl, F. Jelinek, and R. Mercer. "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-5(2), pp. 179–190, 1983.

[3] L. Fissore, P. Laface, G. Micca, and R. Pieraccini, "Lexical Access to Very Large Vocabularies", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37(8):1197–1213, 1989.

[4] L Fissore, P. Laface, G. Micca and R. Pieraccini. "Performance of a Speaker-Independent Continuous Speech Recognizer." in *Speech Recognition and Understanding*, P. Laface and R. De Mori Ed., Springer Verlag, Series F, Vol. 75, pp. 455–479, 1990.

[5] S. Furui. "Unsupervised Speaker Adaptation Method based on Hierarchical Spectral Clustering." *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37(12):1923–1930, 1989.

[6] S. Lerner, B. Mazor. "Telephone Channel Normalization for Automatic Speech Recognition." ICASSP, pp. I-261, I-264, 1992.

[7] Y. Linde, A. Buzo, and R.M. Gray. "An Algorithm for Vector Quantizer Design," *IEEE Trans. on Communications*, Vol.28, pp. 88–95, 1980.