FAST CHANNEL AND NOISE COMPENSATION IN THE SPECTRAL DOMAIN

Christophe Cerisara, Dominique Fohr
LORIA UMR 7503, BP 239 - F54506 Vandoeuvre, FRANCE
Tel: +33(0)383593071; fax: +33(0)383278319
e-mail: cerisara,fohr@loria.fr

ABSTRACT

We compare in this work several methods for fast adaptation of speech models to convolutional and additive noise. The tested algorithms are Parallel Model Combination (PMC), Cepstral Mean Subtraction (CMS), and an algorithm that combines PMC and CMS in the spectral domain. Experiments are realized on a natural numbers recognition task in French. We have trained the acoustic models on the SPEECHDAT database (recorded through telephone lines), and we have tested the system on the VODIS database (recorded in three different cars).

1 INTRODUCTION

Most nowadays automatic speech recognition systems use some kind of model adaptation to deal with mismatches between the training and test environments. These adaptation algorithms usually adapts the models either to a new speaker, or to the background noise. In this work, we consider the latter case, that is called environment adaptation. We assume that two types of noise may corrupt the incoming speech signal: additive noise in the spectral domain, like street or engine noise, and convolutional noise, sometimes called channel noise, which can be modeled by a filter in the spectral domain, like microphone frequency responses.

To adapt the speech models to a new environment, a model of the additive and channel noise has to be estimated using some of the incoming signal. We are only interested here to fast adaptation schemes, i.e. adaptation algorithms that require very few signal to estimate the noise characteristics, and that further have a very low complexity, in order to be able to adapt the speech models in real time every time a pause is detected. Such fast algorithms can be used as bootstrap methods for more complex adaptation algorithms.

To reduce the cost of adaptation, we consider in this work methods that adapt the static cepstral mean vectors of the Gaussian mixtures only.

Section 2 briefly presents the methods that are related to the algorithm proposed in section 3. Section 4 presents the experimental results of several adaptation methods, and section 5 concludes the paper.

2 FAST ADAPTATION METHODS

2.1 Additive noise adaptation methods

One of the most famous method for additive noise adaptation is Parallel Model Combination (PMC). The basics of this method are described in [1]. As we are primary concerned with *fast* adaptation methods, we use in this work simplifying assumptions for PMC. Explicitly, we adapt only the static mean cepstral vectors of the Gaussian mixtures, and we further model the background noise with a single centroid. This simple noise model can be estimated using less than 100 ms of signal.

Several other fast adaptation methods, based on linear approximation of PMC in the cepstral domain, such as VTS [2] or Jacobian adaptation [3], exist.

2.2 Convolutional noise adaptation methods

We consider here only the most widely used algorithm for convolutional noise adaptation, that is Cepstral Mean Subtraction (CMS) [4]. It simply consists to subtract the cepstral average of the signal and of the models.

2.3 Environment adaptation methods

Several methods for additive noise compensation have been extended to also compensate for convolutional noise, such as PMC in [5] or Jacobian adaptation in [6]. Some other algorithms have also been proposed to explicitly combine two adaptation methods, for example PMC and CMS in [7].

3 ADAPTATION IN THE SPECTRAL DOMAIN

3.1 Motivations

CMS is often used because it is an easy-to-implement solution to adapt the system for microphone mismatches. It is easy to implement as it is realized in the cepstral domain, which is the domain of the speech models. However, it is also known to provide a poor adaptation

scheme when both additive and convolutional noise occur.

Originally, the goal of CMS was to match the frequency response of the training and test database, and it was realized in the spectral domain [8]. As PMC anyhow transforms the models into the spectral domain for adaptation, we propose to combine PMC and a channel-matching algorithm, which is directly derived from CMS, in the spectral domain. The algorithm is described in next section.

3.2 Algorithm

3.2.1 Notations

In the next sections, we use the following notations:

- The incoming signal is transformed into the sequence $(Y(t))_{1\leq t\leq T}$ of T frames. Each $Y(t)$ is a spectral vector. We can decompose each frame using the usual environment equation, in the spectral domain:
$$Y(t) = H^{tar}X(t) + N^{tar}$$
where $X(t)$ is the clean signal (without any noise) and N^{tar} and H^{tar} respectively represent the additive and convolutional noise of the incoming signal.

- Adaptation is only performed on the mean vectors of the Gaussian mixtures of the models. Thus, these Gaussians are represented in the spectral domain by the set of vectors $\{G(n)\}_{1\leq n\leq N}$, where N is the total number of Gaussians. We use the same environment equation for the Gaussians than the signal, i.e.:
$$G(n) = H^{ref}S(n) + N^{ref}$$
where N^{ref} and H^{ref} respectively represent the additive and convolutional noise present in the training database. $S(n)$ represents an "ideal" Gaussian, without any noise nor distortion.

3.2.2 Algorithm

Step 1: Gaussians pre-treatment (Training time)

- N^{ref} is computed from the models, by choosing for example one Gaussian of the central state of the silence model.

- The average of all the Gaussians is computed:
$$\mu_G = \frac{1}{N}\sum_{n=1}^{N} G(n) = \frac{1}{N}\sum_{n=1}^{N}\left(H^{ref}S(n) + N^{ref}\right)$$

- N^{ref} is subtracted from this mean:
$$\mu_S = \mu_G - N^{ref} = \left(\frac{1}{N}\sum_{n=1}^{N} S(n)\right) H^{ref} \qquad (1)$$

Step 2: Signal pre-treatment (Test time)

- N^{tar} is computed from the signal, for example by averaging the 10 first frames of each sentence.

- The average of all the signal frames is computed:
$$\mu_Y = \frac{1}{T}\sum_{t=1}^{T} Y(t) = \frac{1}{T}\sum_{t=1}^{T}\left(H^{tar}X(t) + N^{tar}\right)$$

- N^{tar} is subtracted from this mean:
$$\mu_X = \mu_Y - N^{tar} = \left(\frac{1}{T}\sum_{t=1}^{T} X(t)\right) H^{tar} \qquad (2)$$

Step 3: Computation of the channel adaptation factor

- By assuming that
$$\frac{1}{N}\sum_{n=1}^{N} S(n) \simeq \frac{1}{T}\sum_{t=1}^{T} X(t) \qquad (3)$$
the ratio of the two previous averages gives:
$$k = \frac{\mu_X}{\mu_S} = \frac{H^{tar}}{H^{ref}}$$

- It is possible to use a weaker assumption, when the state-frame alignment is known. Let us call $\phi(t) = n$ the index of the Gaussian $G(n)$ aligned with the signal frame $Y(t)$. Then, we can compute and assume that
$$\frac{1}{T}\sum_{t=1}^{T} S(\phi(t)) \simeq \frac{1}{T}\sum_{t=1}^{T} X(t) \qquad (4)$$
However, in this case, the average of the Gaussians needs to be computed at testing time.

Step 4: Adaptation of the Gaussians

- First, channel adaptation is realized by multiplying each Gaussian by k:
$$G_a(n) = kG(n) = H^{tar}S(n) + \frac{H^{tar}}{H^{ref}}N^{ref}$$

- Similarly, N^{ref} is multiplied by k to obtain
$$N_a^{ref} = kN^{ref} = \frac{H^{tar}}{H^{ref}}N^{ref}$$

- Additive noise compensation is finally applied in the spectral domain by adding the bias $N^{tar} - N_a^{ref}$ to the Gaussians:
$$G_{aa}(n) = G_a(n) + N^{tar} - N_a^{ref} = H^{tar}S(n) + N^{tar}$$

- The adapted Gaussians $G_{aa}(n)$ are then transformed into the cepstral domain and used in the Viterbi algorithm for recognition.

One can remark that this algorithm can not be directly used for real-time adaptation, because of channel adaptation that uses the same paradigm as CMS, and thus computes the average of the signal on the whole sentence before recognition. However, it is really easy to make this algorithm real-time, as it is often done for CMS, for example by averaging the signal on the previous sentence, or by using a frame-synchronous estimate of the signal mean.

3.2.3 Cost of the algorithm

The cost of this algorithm is nearly the same as PMC: indeed, the speech signal in the spectral domain can be obtained from the front-end, before cepstral coding. The Gaussians are also already transformed into the spectral domain by PMC. The only additional costs are due to:

- The averaging of the speech signal;
- The multiplication of the Gaussians by k.

Both these costs are very reasonable compared to the cepstral transformation carried out by PMC for each Gaussian.

3.3 Comparison

- This algorithm bears resemblances with the algorithm proposed in [7]. However, there are two majors differences:
  - [7] combines PMC and the usual version of CMS, that is realized in the cepstral domain;
  - [7] assumes that the initial models have been trained in a clean environment ($N^{ref} = 0$).
- Compared to PMC for channel and additive noise compensation, the main difference is again the fact that the channel mismatch is estimated in our case in the spectral domain. We believe the algorithm described in section 3.2.2 can be considered as a simplified implementation of PMC, designed for fast adaptation.

4 EXPERIMENTS

4.1 Experimental setup

4.1.1 Databases

In real applications, speech recognition systems might frequently be built on a given database and used with another microphone and in different noise conditions. To simulate this situation, we have trained our speech models on the SPEECHDAT database, that has been recorded through telephone, and tested it on the VODIS database, that has been recorded in a car environment. The SPEECHDAT database has been realized by recording 1017 speakers through their personal phone. The VODIS database has been recorded using a far-talking microphone, fixed at the rear-view mirror inside three different cars. The speaker is simultaneously talking and driving the car.

The mismatch between the two databases is thus very important and realistic, as it corresponds to the most frequent situations in which speech recognition systems are needed, namely in cars and through telephone.

The task consists to recognize an unconstrained sequence of natural numbers in the French language. A simple loop grammar without bigram probabilities allows to pronounce any sequence of numbers, even meaningless.

4.1.2 Models and system

We have used the HTK Toolkit [9] to generate the 29 speech models. These are standard HMMs with 13 emitting states for the numbers and 3 emitting states for the silence. We have used 8 mixtures per state with diagonal covariance matrices. The signal is encoded into vectors of 13 MFCC coefficients, plus first and second-order derivatives.

4.1.3 Implementation details

We have used a two-pass recognition algorithm:

- In the first pass, the system extracts the 10 frames with the lowest energy to estimate the target noise, and also computes the average of the cepstral vectors for channel adaptation.
- The models are adapted at the end of the first pass, and recognition is realized in the second pass with the Viterbi algorithm.

We have also implemented our algorithm with three passes, to test the weaker assumption proposed in equation 4. In this case, we add to the preceding system a third pass to re-adapt the models using the same algorithm as in the second pass, but with equation 4 instead of 3, based on the alignment obtained in the second pass. In equation 4, the summation is realized only for the *speech* Gaussians that have been aligned with a frame.

4.1.4 Experiments

We have realized two sets of experiments: the first one on the original VODIS database, and the second one on the same database, but filtered through a high-pass filter with a cut-off frequency equal to 1.5 kHz and an attenuation of 5 dB in the low frequencies. Filtering has been applied to emphasize the effect of convolutional noise and observe the adaptation algorithms in such conditions. However, such an important channel mismatch is very rare in practice, and the first tests set gives a better idea of the behavior of the adaptation algorithms.

Figure 1 represents the average frequency responses, i.e. the function γH^{ref} and γH^{tar} with γ a constant term, for the two databases SPEECHDAT and VODIS.

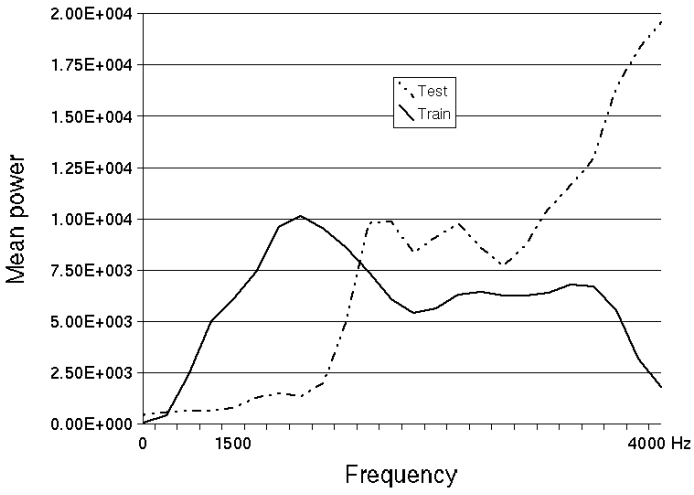

Figure 1: Channel mismatch between the training and test corpus: the two curves are γH^{ref} and γH^{tar}.

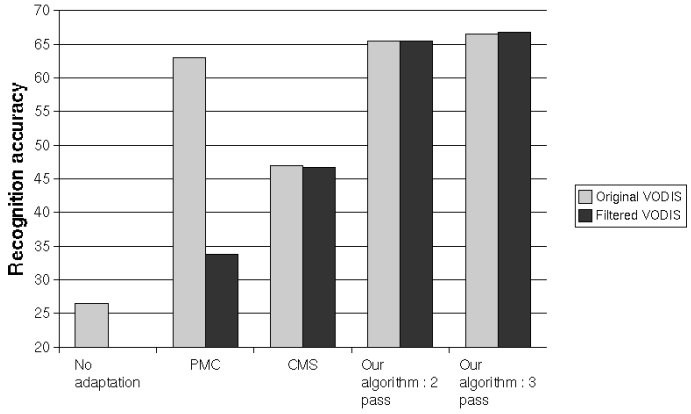

Figure 2: Experimental results

4.2 Experimental results

Several remarks can be made about figure 2:

- Without adaptation, the recognition rates are very poor, which shows that the mismatch between the two databases is very important.
- The best systems are the ones that combine channel and noise compensation. However, we can note that PMC is not far below, which shows that it is a very effective adaptation method, at least when the convolutional noise is not too much important.
- When the VODIS database is filtered, PMC accuracy is much lower, which is expected, as the channel mismatch is now very important. On the other hand, the adaptation methods that explicitly take into account the convolutional noise are not impaired by this additional filtering.

5 CONCLUSIONS

We have shown in this work that, in realistic conditions where the mismatch between the training and test databases is important, a good adaptation scheme can be obtained with a simple and fast compensation algorithm that operates in the spectral domain.

Although we have used a two-pass algorithm, it is easy to transform this algorithm into a one-pass algorithm, as it is usually done for Cepstral Mean Subtraction, for example by computing the average of the speech frames on the previous sentence, of by updating an estimation of this average frame by frame.

References

[1] M.J.F. Gales, *Model-Based Techniques For Noise Robust Speech Recognition*, Ph.D. thesis, Gonville and Caius College, September 1995.

[2] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment independent speech recognition," in *ICASSP'96*, 1996, pp. 733–736.

[3] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *ICASSP'97*, Munich, 1997, pp. 835–838.

[4] A. Anastasakos, F. Kubala, J. Makhoul, and R. Schwartz, "Adaptation to new microphones using tied-mixture normalization," in *ICASSP'94*, 1994, pp. 433–437.

[5] M. Gales and S. Young, "Pmc for speech recognition in additive and convolutional noise," Tech. Rep. CUED/F-INFENG/TR154, Cambridge University Engineering Department, Cambridge, England, December 1993.

[6] S. Sagayama, Y. Kato, M. Nakai, and H. Shimodaira, "Jacobian approach to joint adaptation to noise, channel and vocal tract length," in *ISCA Workshop on Adaptation Methods*, Sophia Antipolis, France, Aug. 2001, pp. 117–120.

[7] H. Yamamoto, T. Kosaka, M. Yamada, Y. Komori, and M. Fujita, "Fast speech recognition algorithm under noisy environment using modified cms-pmc and improved idmm+sq," in *ICASSP'97*, Munich, 1997, pp. 847–850.

[8] T. Stockham, T. Cannon, and R. Ingerbretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, pp. 678–692, 1975.

[9] P. Woodland and S. Young, "The htk continuous speech recogniser," in *Eurospeech'93*, Berlin, Sept. 1993, pp. 2207–2219.