

NON-LINEAR ACOUSTICAL PRE-PROCESSING FOR MULTIPLE SAMPLING RATES ASR AND ASR IN NOISY CONDITION

Richard LAMY, Laurent BESACIER

Equipe *GEOD* (Groupe d'Etude sur l'Oral et le Dialogue) - Laboratoire *CLIPS / IMAG*
Université Joseph Fourier - BP 53 - 38041 GRENOBLE Cedex 9
Tél : (33) 04 76 63 55 81 – Fax : (33) 04 76 63 55 52
Richard.Lamy@imag.fr ; Laurent.Besacier@imag.fr

ABSTRACT

This paper presents a non linear approach for acoustical pre-processing based on Vector Quantization. The idea is to transform feature vectors extracted from signals of one quality to feature vectors of another quality. Our method is applied to two particular cases : speech recognition at multiple sampling rates, and speech recognition in noisy environment. Such a method, which could be applied to other adaptation problems, allows very acceptable correspondence between two considered feature spaces. Thus, a generic ASR system trained on 16kHz signals is able to recognize lower sampling rate signals without any adaptation of its acoustic models. In the same way, our method is applied to ASR in noisy environment and its performance is better than conventional MLLR adaptation when large amount of adaptation data is provided.

1. INTRODUCTION

Automatic speech recognition is applied to many signals of different qualities (sampling rate, quantification, coding, recording conditions). To face this variability, we can use an acoustic model learned on high quality data and transform signals that we want to recognize and/or the reference acoustic model to reduce the mismatch between training and test. This transformation can take place at the signal level, at the acoustic parameters level or by transformation of the model itself. In this last category, we find particular adaptation methods like MAP or MLLR which carry out a modification on the acoustic model distributions. Our approach rather lines up in the acoustical pre-processing (i.e. at feature vector level). The principal advantage of this approach is that, generally, the initial recognition system is not modified by the adaptation process. The most current techniques for such an approach are generally based on linear parameter space transformations, like eigenvoices or eigenrooms, for instance [1]. We present a non linear approach here, using the general principle of Vector Quantization [2]. The first

part of this article will be dedicated to the description of this VQ-based method. In sections 3, 4 and 5, we will present experiments and results for two different applications of our method : multiple sampling rate ASR (notably 8kHz to 16kHz signals adaptation) and ASR in noisy environments. Finally section 6 will conclude this article and give some perspectives.

2. VQ-BASED TRANSFORMATION

Transformations operating in cepstral domain were frequently used in the last decade, like the SDCN algorithm [3] which add a compensation vector that depends on the SNR of the input frame. Nevertheless, the more frequent approach uses probabilities and maximum likelihood criterion. The CDCN algorithm [4] uses EM techniques to compute ML estimates of the environment parameters based on a codebook of *a priori* information. However, SDCN and CDCN techniques only act in the neighborhood of the original acoustic vectors since a linear transformation is done by applying a compensation vector. On the contrary, our technique is geometrical and one acoustic vector can be totally transformed from one feature space to another. The VQ-based transformation learns transformation from acoustic vectors in a source quality to acoustic vectors in a target quality. The goal here is to cluster the source acoustic vectors in a certain number of classes, and then to associate a representative vector in the target format for each class. For doing this, a stereo corpus is needed where a same signal is duplicated in both source and target qualities. The use of stereo database is not new since in the SPLICE Algorithm [5] a linear mapping between two cepstral spaces is learned as a mixture of gaussians.

Figure 1 describes the first classification step of our method : first, since all the vector coefficients are not expressed in homogeneous units and do not represent the same characteristics [6] they are centered and reduced. Then, for the classification of these reduced centered vectors, we use the "binary split" algorithm of the k-means

[7] to cluster the vectors and to associate a target average vector to each class. Former comparative tests between several classification algorithms [8] confirm this choice. At the end of this process, we have the centroid and equivalent of each class.

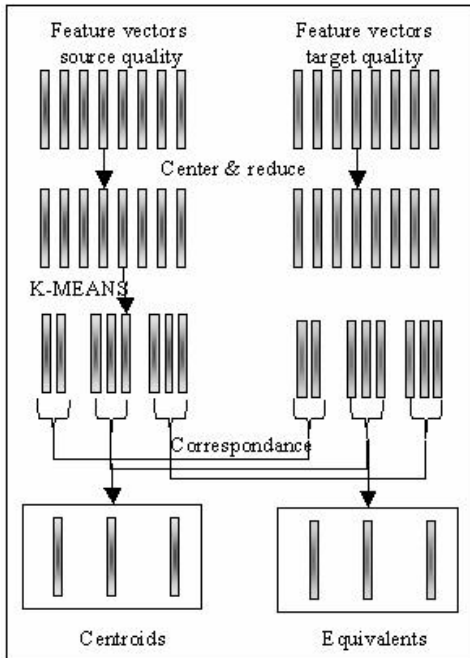


Figure 1 : classification step

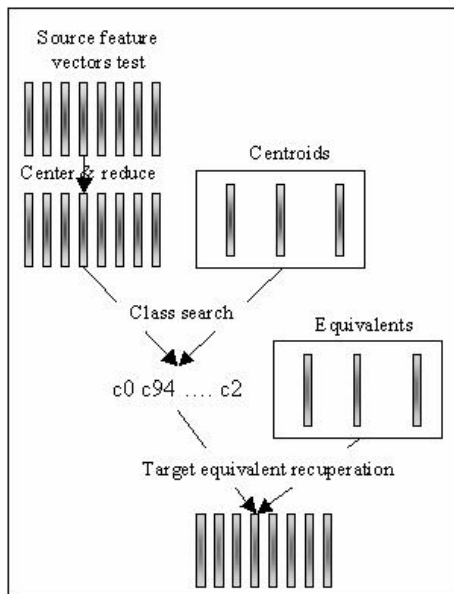


Figure 2 : non-linear pre-processing during recognition

Figure 2 shows the transformation applied to the test signal during recognition. The acoustic vectors of the signal to be recognized are firstly centered and reduced, then compared with centroids to know their belonging class in the source quality. This phase carried out, these vectors are replaced by the equivalent vectors of the same

class but in the target quality space. The recognition phase is done on these new transformed vectors.

3. EXPERIMENTAL FRAMEWORK

We have experimented this method on two different tasks :

-The first one is the adaptation of signals when the sampling rate of the test signals is different of that of the training signals. More precisely, the test signals are sampled at 8kHz and the system is trained on 16kHz signals.

-The second one is the adaptation to noisy signals. Here, the recognition system is trained on clean signals, and we want to recognize noisy signals with it.

For our experiments, we used several English speech "stereos" corpora. Here "stereo" means that the signals of a corpus are duplicated in various qualities and can be aligned vector by vector. Thus, we used two well-known corpus : TIDIGITS [9] and AURORA [10]. Both were split up into three parts : training, adaptation (for calculation of VQ transformations) and test data.

3.1 Ti-digits

TI-DIGITS is a corpus of english connected digits. In order to work on a "stereo" database, we have downsampled this corpus from 16 kHz to 8 kHz. Our training part corresponds to the train part of TIDIGITS (6700 signals). Our adaptation part is composed of 2501 signals of the test part of TIDIGITS, and our test part is made up of the 3450 remaining signals of the test part of TIDIGITS.

3.2 Aurora

AURORA2 database is a corpus of english connected digits in clean and noisy environments. Our training part corresponds to 6440 signals of the train part of AURORA. Our adaptation part is composed of the 2000 remaining signals of the train part of AURORA, and our test part is the "test A" and "test C" parts of AURORA (1001 signals of different noise and different SNR). This cutting is done such that the adaptation part contains 400 signals of each SNR, and 400 signals of each noise. Figure 3 summarizes this partition.

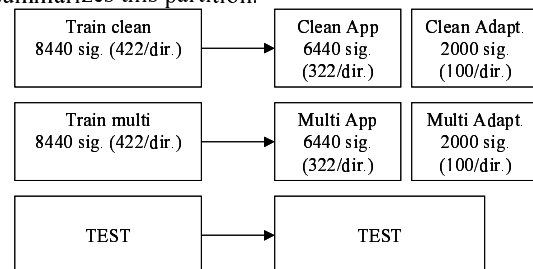


Figure 3 : our AURORA partition

3.3 ASR system

Our english digits recognition system uses Janus-III toolkit [11] from CMU. The phoneme-based acoustic model (3 states HMM, 16 gaussians each) is learned on the train part of respectively each corpus (TIDIGITS or AURORA).

For the feature vectors, we use 13 MFCC coefficients, their first and second derivative, the energy, its first and second derivative, and the zerocrossing rate, all extracted every 10ms on 20ms-windows. These vectors of dimension 43 are then reduced to a dimension of 24 using LDA (Linear Discriminating Analysis). However, when our acoustical pre-processing is used, we do not apply it on the vectors of dimension 43, but only on the static coefficients (MFCC, zerocrossing and energy). The first and second derivative are then recomputed starting from the generated static vector, and the LDA is finally applied.

4. FIRST APPLICATION : MULTIPLE SAMPLING RATE ASR

This first experiment deals with multiple sampling rate ASR, in particular 8kHz to 16kHz adaptation. The results are presented on *table 1*. For comparison purpose, ASR performance obtained when the test 8kHz signals are upsampled to 16kHz, and recognized by 16kHz quality models (81.9%) is given, as well as the performance obtained for 8kHz signals recognized by 8kHz quality acoustic models (96.7%). Our method uses a 13 bits codebook (i.e : 8192 classes) for feature transformation. These results show that our method carries out a good classification of the feature space, in the sense that the sampled 8kHz-sampled signals are well recognized by a recognition system which has never met signals of this quality during training.

quality of signals used to train Acoustic Models	8kHz	16kHz	16kHz
quality of test signals	8kHz	8kHz upsampled to 16kHz	8kHz + N-L pre-processing
word accuracy rate	96.7%	81.9%	95.2%

Table 1 : Performance of our acoustical pre-processing method on English digits

5. SECOND APPLICATION : ASR IN NOISY ENVIRONMENT

For this second experiment, we wanted to compare the effect of our method with a famous adaptation method like MLLR (Maximum Likelihood Linear Regression). We performed two different types of pre-processing : global feature transformation for noisy environment (by using

multicondition data as a whole) and feature transformation specialized to a particular noise type (by using separately all the data of a same noise). These experiments are described in the following subsections.

5.1. Feature transformation with multicondition signals for ASR system trained on clean signals

The results given in this section are the average of all rates obtained for all signals, that is for all SNR levels and all noise types of test A and test C data.

Here again, for comparison purpose, we have two reference scores which are :

- all test A and test C signals recognized by a system trained on clean signals (resp. 54.5% and 55.3%)
- all test A and test C signals recognized by a system trained on multicondition signals (multicondition training of AURORA, resp 73.2% and 66.9%).

We have first applied the MLLR technique to adapt the clean acoustic model in order to obtain a new adapted acoustic model. We obtain both rates given in *table 2* (67.6% for test A and 63.2% for test C). Then, we have computed a correspondence between both feature spaces (clean and noisy) with our VQ-method and sent the reconstructed feature vectors to the original system trained on clean data.

acoustical pre-processing	Acoustic Models	word accuracy rate
NONE	clean	54.5%
NONE	multi	73.2%
NONE	Clean Model adapted by MLLR	67.6%
VQ computed on multi-condition	clean	73.9%

Table 2 : Results on Test A of multicondition adaptation : MLLR adaptation versus VQ pre-processing.

acoustical pre-processing	Acoustic Models	word accuracy rate
NONE	clean	55.3%
NONE	multi	66.9%
NONE	Clean Model adapted by MLLR	63.2%
VQ computed on multi-condition	clean	72.8%

Table 3 : Results on Test C of multicondition adaptation : MLLR adaptation versus VQ pre-processing.

The results show us that the effect of this acoustical pre-processing on the effectiveness of speech recognition is not only better than MLLR adaptation, but also better than the rate obtained with a dedicated acoustic model. This pre-processing gives good results for unknown

environment condition as it is shown by the results for the test C (with noises which have never been seen during the pre-processing or during the acoustic models learning).

5.2. Feature transformation specialized by noise type

This part aims at testing the efficiency of our pre-processing method supposing we have *a priori* knowledge of the environment (i.e. the type of noise). It could then be combined with an environmental quality detection.

The results given in this section are the average rates calculated separately for each noise (N1 subway, N2 babble, N3 car, N4 exhibition) of test A data. This experiment is related to specialized adaptation by noise (N1, N2, N3 and N4), whatever the SNR level. The results presented in *table 4* are WAR obtained on test A part, except on clean signals which contain no noise.

In this experiment, we can see firstly that a specialized MLLR adaptation by noise is surprisingly less efficient than the multicondition training. Secondly, the efficiency of a specific acoustical pre-processing by noise is better not only than the first reference (that is the multicondition training), but also than the pre-processing developed on multicondition corpus. It should be noticed that this pre-processing specialized by noise needs no more development data than the one developed on multicondition (see figure 3) but it however makes the hypothesis that the noise type is *a priori* known.

acoustical pre-processing	acoustic models	N1	N2	N3	N4	AVERAGE
NONE	MultiCondition	72.2%	69.1%	68.9%	68.2%	69.6%
NONE	Clean model adapted by MLLR specialized by noise	69.0%	63.8%	64.4%	65.5%	65.7%
VQ computed on multicondition	Clean	72.4%	69.2%	73.9%	69.0%	71.1%
VQ computed on each noise	Clean	74.0%	70.9%	74.9%	70.7%	72.6%

Table 4 : Word accuracy rates of specialized adaptation for different noises of test A without clean signals.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a non linear approach for acoustical pre-processing. This method, tested on two tasks, gives good results, which are positively comparable with well-known model adaptation techniques like MLLR. We have shown that this VQ-based method allows an efficient transformation of the feature space but it needs a sufficient amount of learning data whereas MLLR is also efficient when few adaptation material is available.

Moreover, it should also be noticed that this technique has a major drawback : it needs a “stereo” corpus, that is exactly the same corpus declined in both qualities (the source one and the target one). We currently work on alignment techniques in order to avoid this drawback.

Another current work is a global recognition system, composed of different transformation objects, which detects the quality of the test signal and applies the correspondent transformation. We can thus recognize many signals without any *a priori* knowledge about it. Our first results present a high rate of quality detection.

7. REFERENCES

- [1] Couvreur L., Dupont S., Ris C., Boite J-M., Couvreur C., “Fast Adaptation for Robust Speech Recognition in Reverberant Environments”, Proc. ITRW 2001 Sophia-Antipolis, pp 85-88, 2001.
- [2] Gersho A., Gray R.M., *Vector quantization and signal compression*, Kluwer Academic Publishers, BOSTON, 1992.
- [3] Acero A., “Acoustical and environmental robustness in automatic speech recognition”, PhD thesis, Carnegie Mellon University, 1990.
- [4] Stern R., Liu F-H., Ohshima Y., Sullivan T.M., Acero A., “Multiple approaches to robust speech recognition”, Proceedings of Speech and Natural Language Workshop, Defense Advanced Research Projects Agency, pp. 274—279, February 1992
- [5] Droppo J., Deng L., Acero A., “Evaluation of the SPLICE Algorithm on the Aurora2 Database”, Eurospeech, 2001, vol. 1, pp. 217—220, 2001.
- [6] Kinnunen T., Kärkkäinen I., Fränti P., “Is speech data clustered ? – statistical analysis of cepstral features”, Eurospeech 2001, vol. 4, pp. 2627-2630, 2001.
- [7] Rabiner, L. & Juang, B-H., “Fundamentals of speech recognition”, pp 126-127., 1993.
- [8] Kinnunen T., Kilpeläinen T., Fränti P., “Comparison of clustering algorithms in speaker identification”, SPC 2000.
- [9] R. G. Leonard. “A database for speaker-independent digit recognition.”, Proceedings of ICASSP 1984, vol. 3, 1984.
- [10] H. G. Hirsch and D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions”, ISCA ITRW ASR2000, Paris, September 2000.
- [11] Soltau H., Schaaf T., Metze F., Waibel A., “the ISL Evaluation System for Verbmobil – II”, in proc ICASSP 2001, salt lake city, USA, May 2001.