LEARNING VECTOR QUANTIZATION AND NEURAL PREDICTIVE CODING FOR NONLINEAR SPEECH FEATURE EXTRACTION

M. Chetouani, B. Gas, J.L. Zarader

Laboratoire des Instruments et Systèmes d'Ile-De-France
Université Paris VI
BP 164, Tour 22-12 2ème étage
4 Place Jussieu, 75252 Paris Cedex 05, France
mohamed.chetouani@lis.jussieu.fr gas@ccr.jussieu.fr zarader@ccr.jussieu.fr

ABSTRACT

Speech recognition is a special field of pattern recognition. In order to improve the performances of the systems, one can opt for several ways and among them the design of a feature extractor. This paper presents a new nonlinear feature extraction method based on the Learning Vector Quantization (LVQ) and the Neural Predictive Coding (NPC). The key idea of this work is to design a feature extractor, the NPC, by the introduction of discriminant constraint provided by the LVQ classifier. The performances are estimated on a phoneme classification task by several methods: GMM, MLP, LVQ. The phonemes are extracted from the NTIMIT database. We make comparisons with linear and nonlinear feature transformation methods (LDA, PCA, NLDA, NPCA), and also with coding methods (LPC, MFCC, PLP).

1. INTRODUCTION

The Neural Predictive Coding (NPC) [3] approach has shown good performances on speech feature extraction. We have developed a discriminant criterion for predictive models: the Modelisation Error Ratio (MER) [2]. This criterion is related to the Maximization of the Mutual Information (MMI). This principle allows the design of a discriminant feature extractor (called the DFE-NPC) in an independent way of the classifier. Indeed, during the DFE-NPC parameterization phase there is no cooperation with the next stage: the classification stage.

Juang [8] have proposed the Discriminative Feature Extraction (DFE) method based on the Minimum Classification Error (MCE) criterion. This criterion focuses on the performances of both the feature extractor and the classifier. The entire recognizer, the feature extractor and the classifier, is trained for a same objective: obtaining the minimum classification error. This is realized by the definition of a loss function that reflects the classification performances of the recognizer. Given this function, a misclassification measure is defined in order to measure the distance between one specific class and the others. Then, the Generalized Probabilistic Descent (GPD) [9] is used to minimize the misclassification distance which is passed through a nonlinear and derivable function like the sigmoid function. Biem [1] applied the DFE method to filter bank design and cepstral liftering.

The key idea of the Discriminative Feature Extraction (DFE) method proposed by Katagiri and al. [9] is that the feature extraction and the classification stage have to be simultaneously trained in order to improve the pattern recognition.

In this paper, we focuse on Discriminative Feature Extraction (DFE) by predictive models and Learning Vector Quantization (LVQ). First, we present the NPC model, then we show how this model can extract nonlinear features by the help of a cooperation with the LVQ. After, we present the experimental conditions and the performances of the system for different phoneme groups. Finally, we give some conclusions and prospects.

2. THE NEURAL PREDICTIVE MODEL

The NPC model is a nonlinear extension of the LPC speech coding. This model is based on neural predictor of the speech waveform. The model differs with the model presented in [3], here the model has two hidden layers.

The NPC is used as speech encoder but only the output layer weights are considered as coding vector or feature vector. For that, the learning phase is realized in two times. First, the parameterization phase consists in the learning of all the weights by the prediction error minimization:

$$Q = \sum_{k=1}^{K} (y_k - \hat{y}_k)^2 = \sum_{k=1}^{K} (y_k - F(\mathbf{y}_k))^2 \tag{1}$$

With y speech signal, $\hat{y}$ predicted speech signal, k the samples index and K the number of samples.

In this phase, only the first layer weights $\mathbf{w}$ which are the NPC encoder parameters are kept. Since the NPC encoder is set up by the parameters defined in the previous phase, the second phase, called the coding phase, consists in the computation of the output layer weights $\mathbf{a}$: the phoneme coding vector.

F is a nonlinear function which is decomposed into two functions $G_{\mathbf{w}_{1,2}}$ ($\mathbf{w}_{1,2}$ first and second layers weights) and $H_{\mathbf{a}}$ ($\mathbf{a}$ output layer weights):

$$F_{\mathbf{w}_{1,2},\mathbf{a}}(\mathbf{y}_k) = H_{\mathbf{a}} \circ G_{\mathbf{w}_{1,2}}(\mathbf{y}_k) \tag{2}$$

With $\hat{y}_k = H_{\mathbf{a}}(\mathbf{z}_k)$ and $\mathbf{z}_k = G_{\mathbf{w}_{1,2}}(\mathbf{y}_k)$.

The NPC weights modification law $\Delta\mathbf{a}^{Pred}$ and $\Delta\mathbf{w}_{1,2}^{Pred}$ are proportional to the gradient of the prediction errors $Q_{NPC} = \sum_i^M Q_i$, with M the number of classes.

One of the NPC objective is the discrimination between the features extracted. For that, we developed a discriminant criterion called the MER [2], but without explicit link with classification phase. In the next section, we propose a cooperation with the LVQ classifier.

3. COOPERATION BETWEEN NPC AND LVQ

The Learning Vector Quantization (LVQ) classifier has been successfully applied on different domains like handwritten recognition [10] or speech recognition [11]. The learning procedure consists in the prototypes adjustment in order to describe optimal class boundaries.

The simultaneous training of NPC and LVQ models can be done by the help of the DFE framework. Indeed, this method allows to introduce discrimination provided by the LVQ classifier into the NPC model. The DFE considers the feature extractor and the classifier as a single module described by $\mathbf{\Phi} = (\mathbf{a}, \mathbf{m})$ ($\mathbf{m}$ being the LVQ classifier prototypes).

3.1 DFE framework

The DFE framework needs the definition of a discriminant function. In the case of the LVQ classifier, the discriminant function of a class is the negative of the minimum distance from the input pattern (the feature vector) to the genuine prototype:

$$g_i(\mathbf{a}) = -\min_{\tau} \|\mathbf{a} - \mathbf{m}_{i,\tau}\|^2 = -\min_{\tau} \mathrm{d}(\mathbf{a}, \mathbf{m}_{i,\tau}) \tag{3}$$

where $\mathrm{d}(\mathbf{a}, \mathbf{m}_{i,\tau})$ is the Euclidean distance between the input pattern $\mathbf{a}$ (feature vector) and a prototype $\mathbf{m}_{i,\tau}$ of the class C_i.

According to the DFE framework, one could define the misclassification measure:

$$\mu_i(\mathbf{a}) = -g_i(\mathbf{a}) + \left[\frac{1}{M-1}\sum_{j \neq i} g_j(\mathbf{a})^{-\psi}\right]^{-\frac{1}{\psi}} \tag{4}$$

where ψ is a positive number. For a large ψ, the misclassification measure becomes:

$$\mu_i(\mathbf{a}) = -g_i(\mathbf{a}) + \overline{g}_i(\mathbf{a}) \tag{5}$$

$\overline{g}_i(\mathbf{a})$ is the competing discriminant function (anti-discriminant function) to the class C_i. This leads to only consider the first incorrect prototype [11]:

$$\overline{g}_i(\mathbf{a}) = \max_{j \neq i} g_j(\mathbf{a}) \tag{6}$$

The misclassification measure $\mu_i(\mathbf{a})$ (4) has to be positive when $\mathbf{a}$ is misclassified and negative if this is not the case:

$$\mu_r(\mathbf{a}) = \mathrm{d}(\mathbf{a}, \mathbf{m}_{i,\tau}) - \mathrm{d}(\mathbf{a}, \mathbf{m}_{j,\upsilon}) \tag{7}$$

where $\mathbf{m}_{i,\tau}$ is the closest prototype of the genuine class while $\mathbf{m}_{j,\upsilon}$ is the closest prototype of the incorrect class.

The next step consists in the definition of MCE loss function which reflects the classification errors:

$$l_i(\mathbf{a}) = l_i(\mu_i) = \frac{1}{1 + e^{-\zeta \mu_i}} \tag{8}$$

The MCE objective function is the following empirical loss:

$$L(\mathbf{a}, \mu) = \sum_{n=1}^{N} \sum_{i=1}^{M} l_i(\mathbf{a}_n) \delta_{C(\mathbf{a}_n) - i} \tag{9}$$

where $C(\mathbf{a}_n)$ is the class membership of the feature vector $\mathbf{a}_n$ and δ is the Kronecker symbol which worths 1 when $C(\mathbf{a}_n) = i$. N is the number of frames and M the number of classes.

The Generalized Probabilistic Descent (GPD) is applied for updating the parameters $\mathbf{\Phi} = (\mathbf{a}, \mathbf{m})$:

$$\begin{aligned}
\mathbf{a}_n &= \mathbf{a}_n - \beta(t) \frac{\partial l_i(\mathbf{a}_n)}{\partial \mathbf{a}_n} \\
\mathbf{m}_{i,\tau} &= \mathbf{m}_{i,\tau} - \alpha(t) \frac{\partial l_i(\mathbf{a}_n)}{\partial \mathbf{m}_{i,\tau}} \\
\mathbf{m}_{j,\upsilon} &= \mathbf{m}_{j,\upsilon} - \alpha(t) \frac{\partial l_i(\mathbf{a}_n)}{\partial \mathbf{m}_{j,\upsilon}}
\end{aligned} \tag{10}$$

where $\alpha(t)$ and $\beta(t)$ are the learning rates of respectively the LVQ classifier and the NPC model. The learning rates are decreasing function of the epoch index t.

According to the MCE loss function (8), the updating rules for the LVQ parameters are as:

$$\begin{aligned}
\mathbf{m}_{i,\tau} &= \mathbf{m}_{i,\tau} + 2\alpha(t) l_i(\mathbf{a}_n)(1 - l_i(\mathbf{a}_n))(\mathbf{a}_n - \mathbf{m}_{i,\tau}) \\
\mathbf{m}_{j,\upsilon} &= \mathbf{m}_{j,\upsilon} - 2\alpha(t) l_i(\mathbf{a}_n)(1 - l_i(\mathbf{a}_n))(\mathbf{a}_n - \mathbf{m}_{j,\upsilon})
\end{aligned} \tag{11}$$

And for the NPC model, the updating rule for the feature vectors $\mathbf{a}_n$ is as:

$$\Delta \mathbf{a}_n^{MCE} = -2\beta(t) l_i(\mathbf{a}_n)(1 - l_i(\mathbf{a}_n))(\mathbf{m}_{i,\tau} - \mathbf{m}_{j,\upsilon}) \tag{12}$$

One can remark that the feature vectors are updated in a fonction of the distance between the two prototypes: the genuine and the incorrect. They are updated in the direction of the maximum separability between these two classes.

These contributions have to be associated with the prediction modifications: $\Delta \mathbf{w}_{1,2}^{Pred}$ and $\Delta \mathbf{a}^{Pred}$.

3.2 Cooperation

Indeed, the objective of this cooperation is to introduce discriminant constraint on the NPC parameterization phase. Several solutions can be used. For instance one can opt for constraint minimization like in [4] where the simultaneous training of classifiers is processed with a Lagrangian formalism. Here, we opt for another approach. The two optimizations are moderated with the help of a coefficient θ. The resultant modification for the feature vectors $\mathbf{a}_n$ is as:

$$\Delta \mathbf{a} = \theta \Delta \mathbf{a}^{Pred} + (1 - \theta) \Delta \mathbf{a}^{MCE} \tag{13}$$

The second phase of the cooperation consists in the modification of the first layers in the maximum class separability direction. However, the relation between the first layers weights $\mathbf{w}_{1,2}$ and the MCE criterion (8) is not direct as for the output layer weights. Considering the objective of the NPC in the cooperation, that reverts bringing closer the features to their adequate prototypes and to move away them from the incorrect prototypes. In other words, the feature vector $\mathbf{a}_{i,n}$ produced by the NPC model for the analysis window $\mathbf{y}_{i,n}$ (belonging to the class C_i) must be close to one of the prototypes $\mathbf{m}_{i,\tau}$.

For that, we introduce a new stage into the NPC model. For the window $\mathbf{y}_{i,n}$, we determine the two modifications necessary for:

- Bringing the feature to the prototype $\mathbf{m}_{i,\tau}$: minimization of the prediction error under the constraint: the output layer is fixed to $\mathbf{m}_{i,\tau}$. One obtains the modification of the first layers $\Delta \mathbf{w}_{1,2}^{mod}$.
- Move away from the prototype $\mathbf{m}_{j,\upsilon}$: maximization of the prediction error under the constraint: the output layer is fixed to $\mathbf{m}_{j,\upsilon}$. One obtains the modification of the first layers $\Delta \mathbf{w}_{1,2}^{disc}$.

During these two processes, one estimates the modifications necessary to maximize the separability of the classes.

The modification law of the first layers is a moderation of these two effects:

$$\Delta \mathbf{w}_{1,2} = \theta \Delta \mathbf{w}_{1,2}^{mod} + (1 - \theta) \Delta \mathbf{w}_{1,2}^{disc} \tag{14}$$

One can notice that this modification law does not integrate the modification of model NPC $\Delta \mathbf{w}_{1,2}^{Pred}$. Indeed, this modification is not useful any more because the contribution $\Delta \mathbf{w}_{1,2}^{mod}$ makes it possible to take account of the modeling part necessary to the LVQ-NPC process.

4. EXPERIMENTAL CONDITIONS

4.1 Database

The NTIMIT database [7] is used in the experiment of this work. This database is composed by 10 sentences pronounced by 630 speakers of 8 areas of the United States. In using this database, we carry out speech recognition in telephone quality. In this work, we focused in the processing of front vowels (/ih/, /ey/, /eh/, /ae/), voiced plosives (/b/, /d/, /g/) and unvoiced plosives (/p/, /t/, /k/) from the first region DR1 (New England). This choice can be justified by the fact that the classification of these phonemes is known to be difficult and are often used . We used the configuration proposed in the NTIMIT database for the training (24 male and 14 female) and the

Phoneme	ih	ey	eh	ae	b	d	g
Train	316	189	297	346	183	300	167
Test	104	54	83	79	59	90	44
Phoneme	p	t	k				
Train	215	320	390				
Test	48	93	103				

Table 1: Database composition for the training and test phases

test set (7 male and 4 female)). The number of phonemes for each class is described in the table 1

Depending on their duration, each phoneme is split into a number of frames. The analysis window size is fixed to 128 samples (8kHz for NTIMIT) with an overlapping of 64 samples.

4.2 Classification

The proposed work is an evaluation of a new feature extractor: the LVQ-NPC. Consequently, we make comparisons with the most used methods: the Linear Predictive Coding (LPC), the Mel Frequency Cepstral Coding (MFCC), the Perceptual Linear Predictive (PLP) [6] speech coding methods. The feature vector dimension is set to 12.

An efficient way for parameters evaluation is the classification. The classification is done frame by frame without context dependency. Moreover, it must be carrying out by several types of classifier in order to measure the discriminant capacities of each feature extraction method.

4.2.1 Gaussians Mixture Models GMM

This model is based on densities estimation of each class. GMMs are trained by the help of the EM algorithm (Expectation-Maximization) with diagonal matrices of covariance assumption. This classifier is sensitive to initialization, the parameters are initialized by the k-means algorithm (10 iterations) with $k = 16$. Classification is done according to the maximum likelihood (ML) criterion.

4.2.2 Prototypes classification (LVQ)

The LVQ model (Learning Vector Quantization) is a prototype-based classifier. The training and the test are carried put by the consideration of the Euclidean distance. This method is also sensitive to initialization, we also use the k-means with $k = 50$.

4.2.3 Neural networks

The neural networks are based on nonlinear discriminant functions. The model has one hidden layer of 10 neurons and the training is done by the Levenberg-Marquardt algorithm.

4.2.4 LVQ-NPC

The number of prototypes for the LVQ-NPC model is fixed at 25. The moderating parameter θ follows a decreasing law:

$$\theta(t) = \theta_0 \left(1 - \frac{t}{N}\right) \tag{15}$$

where N is the iteration number. The value of θ_0 differs according to the treated phonetic group: 0.6 for the vowels, 0.7 for the plosives (voiced and unvoiced).

Such evolution law shows that it is necessary by starting to model the classes then to increase discrimination progressively.

4.3 Feature reduction

The main objective of the feature extraction step is to extract relevant information directly from speech signals. Usually, the next step in a speech recognition system, is the transformation of these features of dimension p to a lower dimension m. The role of this step is dimension reduction but also it can improve the classification rates. As it is noted in [12], if the feature extractor is properly designed, there is no need for this feature transformation.

Here, we used the feature transformation in order to measure the efficiency of each feature extractor. The transformations used are linear and nonlinear. The Linear Discriminant Analysis (LDA) criterion is based on discriminant separability [5], the Principal Component Analysis (PCA) is based on the maximization of the variations of the original feature space. We also used the nonlinear equivalent transforms by neural networks: NLDA and NPCA. The classification is done by GMMs.

5. PHONEME CLASSIFICATION RESULTS

In this section, we present the results in phoneme classification. The shown classification rates are those of the test.

Table 2 shows the classification rates for the vowels with the various methods of coding and classification. The first remark that one can make is that the proposed coding method, the LVQ-NPC, allows an improvement in the classification scores. The introduction of a nonlinear modeling by neural networks and discrimination allow an improvement of 4% compared to the MFCC coding.

	LPC	MFCC	PLP	LVQ-NPC
GMM	34.94	47.07	43.05	51.32
LVQ	35.27	42.55	39.71	50.89
MLP	40.30	45.39	42.57	50.49

Table 2: Classification rates for the vowels

Due to the fact that those plosives (/b/,/d/,/g/) are voiced phonemes, we find a similar behavior for the LVQ-NPC method (cf. tab. 3).

	LPC	MFCC	PLP	LVQ-NPC
GMM	53.52	58.05	57	65.22
LVQ	51.01	56.26	56.23	63.67
MLP	54.58	55.19	57.07	62.73

Table 3: Classification rates for the voiced plosives

The classification of unvoiced plosives (/p/,/t/,/k/) is interesting because they are unvoiced phonemes, which initially seems to be penalizing for predictive models like the LVQ-NPC. However, discrimination makes it possible to overcome this problem and to obtain better results independently of classification method (cf tab. 4).

	LPC	MFCC	PLP	LVQ-NPC
GMM	43.21	49.52	46.12	51.15
LVQ	42.93	47.16	45.83	49.76
MLP	44.49	47.67	45.52	47.16

Table 4: Classification rates for the unvoiced plosives

The feature reduction process is carried out for the voiced plosives (/b/,/d/,/g/). For each coding vector, we computed the Δ and $\Delta\Delta$ parameters, the resulting vector dimension is 36. We performed dimension reduction with LDA, PCA, NLDA and NPCA.

One can see on the figures (1,2, 3, 4) that the classification rates are improved by feature transformations. The PCA and LDA performances are similar for high dimensions but very different for low dimensions. With the both feature transformation methods, the LVQ-NPC performances are better on all the dimensions. These results show that the LVQ-NPC is not affected by feature reduction. The nonlinear feature transformations allow an improvement of classification rates for all the coding methods: there are nonlinear informations between the feature vector and the Δ and $\Delta\Delta$ parameters. Moreover, one can see that for the LVQ-NPC method (cf.

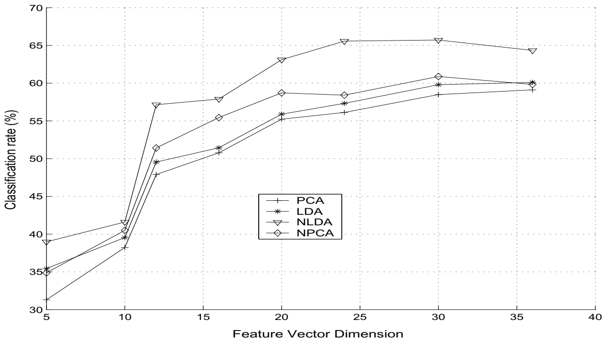

Figure 1: LPC Feature Reduction: results of LDA, PCA, NLDA, NPCA

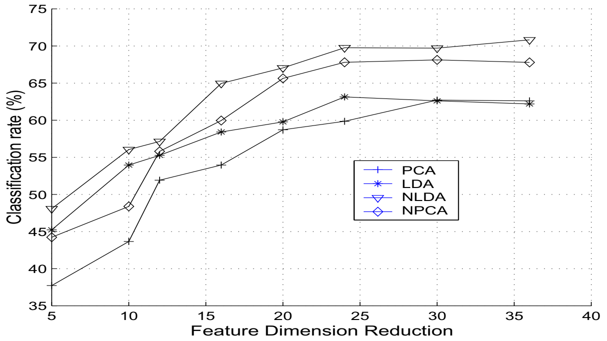

Figure 2: MFCC Feature Reduction: results of LDA, PCA, NLDA, NPCA

figure 4) the improvement is more important than in the other methods. Indeed, this method is a nonlinear feature extraction method, and the appropriated transformation methods should be nonlinear in order to preserve the nonlinear modelization.

6. CONCLUSIONS

We have presented a new coding method: the LVQ-NPC which is based on the simultaneous training of an feature extractor and a classifier. The model is adapted for speech processing which seems required nonlinear modeling but also an adequate discrimination. The experimental results in phoneme classification resulting from NTIMIT database shows the interest of the method. Moreover, the performances measured by classifiers with different behaviors show an improvement in comparison with the traditional coding methods. The features extracted are also robust for dimension reduction by linear and nonlinear methods. We show that for nonlinear feature extraction methods, the appropriated methods are also nonlinear. Our next work consists of a validation on a greater number of phonemes but also on the cooperation with other classifiers.

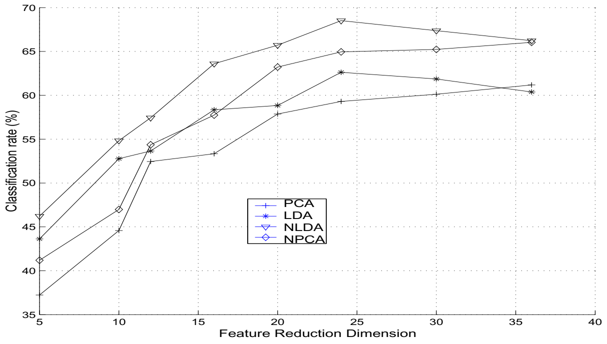

Figure 3: PLP Feature Reduction: results of LDA, PCA, NLDA, NPCA

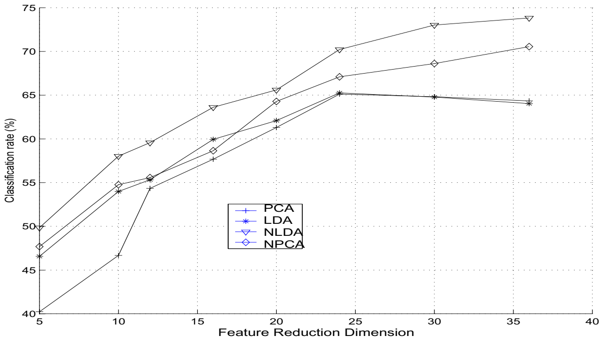

Figure 4: LVQ-NPC Feature Reduction: results of LDA, PCA, NLDA, NPCA

REFERENCES

[1] A. E. Biem. *Discriminative Feature Extraction Applied to Speech Recognition*. PhD thesis, Paris 6, 1997.

[2] M. Chetouani, B. Gas, and J.L. Zarader. Maximization of the modelisation error ratio for neural predictive coding. *Proc. of NOLISP*, 2003.

[3] M. Chetouani, B. Gas, and J.L. Zarader. Modular neural predictive coding for discriminative feature extraction. *Proc. of ICASSP*, 2003.

[4] X. Driancourt. *Optimisation par descente de gradient stochastique de systèmes modulaires combinant réseaux de neurones et programmation dynamique*. PhD thesis, Université Paris XI Orsay, 1994.

[5] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2001.

[6] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, pages 1738–1752, 1990.

[7] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. Ntimit: A phonetically balanced, continous speech, telephone bandwidth speech database. *ICASSP*, 1:109–112, 1990.

[8] B-H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. Signal Processing*, 40:3043–3054, December 1992.

[9] S. Katagiri. *Handbook of Neural Networks for Speech Processing*. Artech House eds., 2000.

[10] C.-L. Liu and M. Nakagawa. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition. *Pattern Recognition*, 34:601–615, 2001.

[11] E. McDermott. *Discriminative Training for Speech Recognition*. PhD thesis, Waseda University (Japan), 1997.

[12] X. Wang and K. D. O'Shaughnessy. Improving the effeciency of automatic speeach recgnition by feature transformation and dimensionality reduction. *EUROSPEECH*, 2003.