IMPROVED ROBUST FEATURES FOR SPEECH RECOGNITION BY INTEGRATING TIME-FREQUENCY PRINCIPAL COMPONENTS (TFPC) AND HISTOGRAM EQUALIZATION (HEQ)

Shang-nien Tsai and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan, Republic of China.
sntsai@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

Robustness for speech recognition technologies with respect to adverse environments has been a key issue for real applications. Time-frequency principal components (TFPC) features were shown to be a set of powerful data-driven features under matched circumstances, while histogram equalization (HEQ) was proposed as an efficient feature transformation approach to reduce the mismatch between training and testing conditions. In this paper, it is proposed that TFPC features can be well integrated with HEQ. HEQ generates a well-matched environment, in which TFPC features can be properly utilized. Extensive experiments with respect to the AURORA2 database verified that improved performance in adverse circumstances can be achieved.

1. INTRODUCTION

The blueprint for the various applications of the automatic speech recognition (ASR) technologies in the future has been extensively laid out and its realization has been highly anticipated by many people [1]. But the recognition accuracy always plays the most dominating role when the realization of real-world applications is considered. It is well known that the recognition accuracy of ASR systems is very often seriously degraded by the mismatch between the acoustic conditions for the training and testing environments and, hence, robustness for ASR technologies with respect to the acoustic environment has always been a key issue in real applications.

One direction towards the above goal is to find some new noise resistant features for speech recognition. Cepstrum mean subtraction (CMS) and cepstrum normalization (CN) are two widely adopted transformations to produce relatively robust features for this purpose due to the relatively low computational requirements as well as the significant achievable improvements [2, 3]. Recently, a new approach for feature transformation based on the concept of histogram equalization (HEQ), which has been widely used in image processing for its capabilities in contrast enhancement, was introduced to speech recognition problems and appears to be superior to the conventional CMS and CN approaches [4].

In most cases, the transformation techniques mentioned above have been applied to Mel frequency cepstral coefficients (MFCC), which have shown consistently satisfactory performance over a wide variety of application tasks. But the optimality for a specific application task is not guaranteed. This is why many new approaches to develop data-driven feature sets have been reported in recent years with a goal that better performance can be obtained for specific application tasks. The time-frequency principal components (TFPC) features represent a good example along this line, which have been shown to provide better performance in some speaker identification and speech recognition tasks than MFCC, but under matched conditions [5, 6]. In this paper we proposed to integrate the TFPC features with the HEQ transformation to produce a new set of features. Since the HEQ may equalize the mismatched conditions, while TFPC features may offer better performance when the mismatched conditions are equalized. Experimental results with respect to the AURORA2 database verified that improved performance in the adverse circumstances can be actually achieved.

The remainder of this paper consists of 4 sections. TFPC and HEQ are very briefly reviewed in section 2, the experimental conditions described in section 3 and extensive experimental results are presented in sections 4. Section 5 finally gives the concluding remarks..

2. TIME-FREQUENCY PRINCIPAL COMPONENTS (TFPC) AND HISTOGRAM EQUALIZATION (HEQ)

Here we briefly summarize the concepts of time-frequency principal components (TFPC) and histogram equalization (HEQ) for development purposes.

2.1. Time-frequency principal components (TFPC)

Because MFCC are not necessarily optimal for a specific application task, many data-driven feature extraction techniques were developed, for example via principal component analysis (PCA), linear discriminant analysis (LDA), or minimum classification error (MCE) criteria. Many of such techniques utilize the concept of temporal filtering, or filtering the temporal trajectories of the MFCC. In other words, these techniques try to include a larger temporal span of the speech features in some way optimized with a corpus characterizing the application task [7, 8].

Compared to the LDA and MCE, those approaches based on PCA have the advantage that no transcription of training data

0-7803-7980-2/03/$20.00 © 2003 IEEE

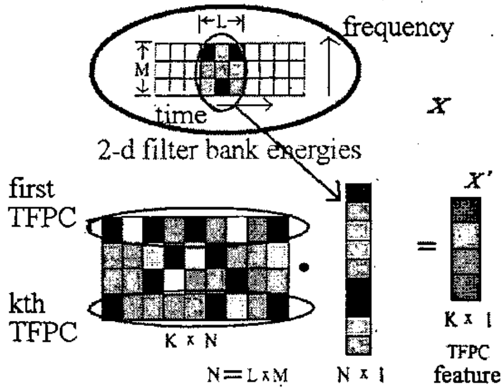


Figure 1: The concept of TFPC feature extraction.

for classification purposes is required, although better or similar performance can still be obtained [8, 9]. Recently PCA was further applied to the design of 2-dimensional data-driven filters to obtain the time-frequency principal components (TFPC) and very promising results were found under matched conditions in speaker identification and speech recognition [5, 6].

PCA is usually used for dimension reduction and decorrelation of feature coefficients. In this technique, K eigenvectors corresponding to the K largest eigenvalues of the covariance matrix of the original observation vectors are taken as the principal components, usually K << N, where N is the dimension of the original observation vectors. The projections of the original N-dimensional observation vectors onto the K eigenvectors are then taken as the K-dimensional new features. It can be shown that any two new feature dimensions thus obtained are statistically independent, and such a transformation is optimal in the MMSE sense [11].

In most of the PCA-based temporal filtering approaches, segments of the temporal trajectories of the same spectral or cepstral parameters (e.g. MFCC) are taken as the original observation vectors. In the TFPC technique proposed recently and used here, on the other hand, a matrix of logarithms of filter bank energies for a total of M different spectral parameters and L temporal frames, which gives M x L = N parameters, is used to construct the original N-dimensional observation vector, $\boldsymbol{X}$. PCA performed on the covariance matrix of a large ensemble of the observation vectors $\boldsymbol{X}$ then produces a total of K N-dimensional eigenvectors, or TFPC, which can be used to transform all the N-dimensional observation vectors $\boldsymbol{X}$ into K-dimensional new feature vectors $\boldsymbol{X}'$, or the TFPC features. This process is shown in Figure 1. In this way, the temporal and spectral variations of the speech signals can be jointly considered in the new TFPC filtering. This process can be repeated when the M x L processing matrix is shifted on the 2-dimensional filter bank energies frame by frame. This can be regarded as finding a "globally", i.e. both temporally and spectrally, optimal independent feature set as an alternative to MFCC.

2.2. Histogram equalization (HEQ)

To relieve the performance degradation due to mismatch in acoustic conditions between training and testing data, including those caused by convolutional and additive noise, cepstrum mean subtraction (CMS) and cepstrum normalization (CN) have been proposed and found very helpful [2, 3]. Recently a new approach of histogram equalization (HEQ) was proposed, which seems to offer even better transformation than CMS and CN [4].

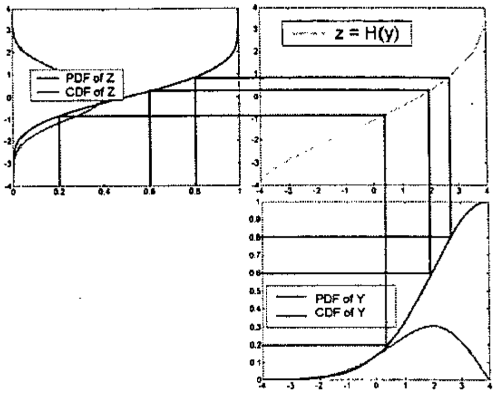


Figure 2: The concept of HEQ.

CMS subtracts from each feature coefficient in a temporal span its mean evaluated over this span, in order to eliminate the dc offset caused by different environments, primarily the convolutional noise. The span can be an utterance or a fixed interval of time [12]. By properly selecting the span as well as its length, higher recognition accuracy can be achieved. CN moves one step further than CMS. It not only removes the dc offset, but also constrains the dynamic range of the features within a certain limit in the temporal span. This is done by first subtracting the mean, and then properly scaling the feature coefficients so that each feature coefficient has a fixed variance. With this additional operation, the mismatch is further alleviated, and therefore causes relatively less harm to the achievable performance.

The recently proposed histogram equalization (HEQ) moves still one step further than CN. It tries to reshape distributions of both the training and testing features in each temporal span (primarily utterance) into standard normal distributions by non-linearly transforming the cumulative histograms of the feature coefficients [4]. The basic idea of HEQ is to define a transformation **H(•)** for a random variable **Y** such that **H(Y)** is a standard normal random variable **Z**, but preserves the cumulative distribution of **Y**, i.e., if

$$\mathbf{Prob\ (\ Y > y\ ) = Prob\ (\ Z > z\ )}$$

,then

$$\mathbf{H(y) = z}$$

This is illustrated in Figure 2. Now, a certain feature coefficient in the training and testing corpora can be taken as two random variables **Y** and **W**, respectively. Due to the mismatched conditions, **Y** and **W** have different distributions. The HEQ discussed here can then force them both become standard normal, which certainly reduces the degree of mismatch.

It should be noticed that the noise effects are usually linear in time and spectrum domains, but becomes non-linear in the feature domain due to some non-linear process in feature extraction, such as taking logarithm. The HEQ process reviewed here is a non-linear operation, therefore may offer a better compensation for the non-linear effect caused by noise than linear operations like CMS or CN does [4].

With the identical cumulative histograms, the transformed training and testing features definitely share the same mean and variance. Yet they also share the same probability density function in addition. Therefore HEQ can be regarded as a stronger transformation which possesses the merits of CMS and CN, but with extra nice properties.

All these feature transformation methods, CMS, CN, and HEQ have shown their effectiveness with MFCC under mismatched conditions, while TFPC features have shown better performance than MFCC under matched conditions. Therefore below in this paper we try to apply these transformations to TFPC features, to find out whether better performance is achievable under adverse conditions.

3. EXPERIMENTAL CONDITIONS

The experiments reported in this paper were conducted on the database AURORA2. 10 different types of noises, as representatives of real-world noises, were included in this database. There are two sets of training conditions in AURORA2 tasks, the clean-speech training and multi-condition training, and there are three sets of testing conditions, each with different mismatched environments, i.e., set A (subway, babble, car, and exhibition noises), set B (restaurant, street, airport, and train station noises), and set C (subway and street noises, with channel effect). The MFCC were obtained using the AURORA2 WI007 Front-end, which gives 13 coefficients (C1 ~ C12 + log energy). The clean speech data in AURORA2 consists of 8440 utterances of English connected digit strings, which were used to train the K x N 2-dimensional parameters as shown in Figure 1 to obtain the TFPC features. In this procedure, PCA was applied to the logarithms of 23 Mel scale filter bank energies with temporal span of length 5 centered at the current time instant, i.e. M=23, L=5, and N= M x L=115 as in Figure 1. So 115-dimensional feature vectors X as in Figure 1 were generated. The number of TFPC, K in Figure 1, is chosen to be 39, identical to that of MFCC with first and second derivatives, for fair comparison.

4. EXPERIMENTAL RESULTS

4.1. Projections of the MFCC bases onto the subspace spanned by TFPC

It can be noted that all the 39 MFCC (static, first and second derivatives) can be regarded as the results of some linear operations performed on the 2-dimensional M x L=115 log energies in Figure 1, if we set the configuration parameters *DELTAWINDOW* and *ACCWINDOW* defined in HTK documentation both to 1. Therefore these 39 MFCC can also be considered as the inner products of the 115-dimensional vector X in Figure 1 with 39 specially defined vectors. These 39 specially defined vectors are thus referred to as MFCC bases here. On the other hand, the 39 eigenvectors produced by PCA performed on the AURORA2 database are referred to as TFPC [5].

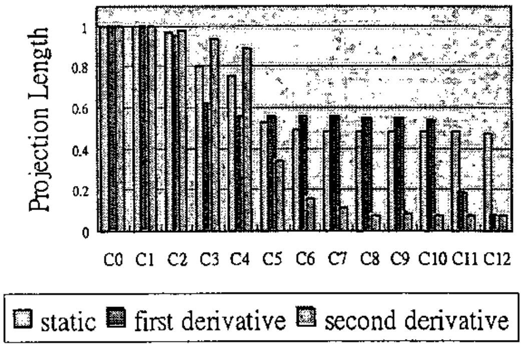


Figure 3: Projection lengths of MFCC bases onto the subspace spanned by TFPC bases.

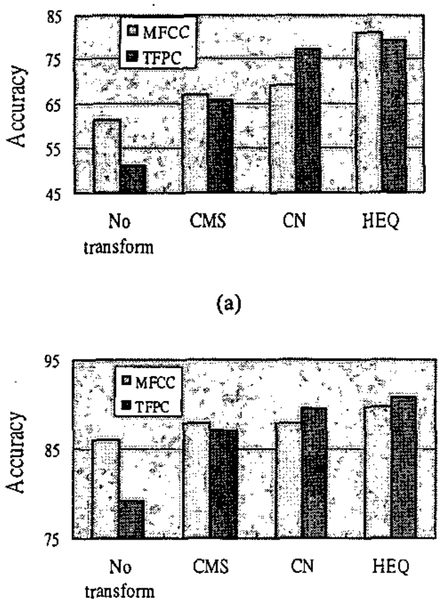
With the above definition of MFCC bases and TFPC, we can then first normalize the MFCC bases into unit length and then project them onto the subspace spanned by the 39 TFPC, and the projection lengths are plotted in Figure 3. Here a projection of length of unity (e.g. the 6 MFCC bases for C0 and C1 in Figure 3) indicates the whole MFCC basis is within the 39 TFPC dimensions, while a projection with length less than unity indicates parts of the MFCC basis are within the 39 TFPC dimensions, while some other parts are not.

From Figure 3, we can see that most of the 39 MFCC bases are more or less reasonably represented in the space spanned by the TFPC, and the bases corresponding to low-order MFCC (e.g. C0 ~ C4), which were known to play key roles in the recognition process [13], are almost completely included in the 39 TFPC dimensions. This implies the most important information for recognition conveyed by MFCC has been very well preserved in the TFPC features with little loss.

4.2. Performance of TFPC integrated with HEQ

In the recognition experiments, after the TFPC features are extracted, they are further processed with CMS, CN, and HEQ, and then used in training the models and performing the recognition. As baseline experiments, the static MFCC are also extracted, processed with CMS, CN, and HEQ, and then used to calculate the derivatives and perform the recognition experiments. The accuracies averaged for the three testing conditions (Sets A, B, and C.) with different noise types and all levels of SNR's (i.e. 0 ~ 20 dB) respectively for clean-speech training and multi-condition training are plotted in Figure 4(a) and (b).

In figure 4(a) and (b), we can see that TFPC features alone, without any transformation applied, are actually much less robust than MFCC under mismatched conditions, with either the clean-speech training or the multi-condition training. In fact, the averaged performance of TFPC in this case is significantly worse than MFCC. It should be noticed that this is not inconsistent with previous results that TFPC features work



(a)

(b)

Figure 4: Average word accuracy for MFCC and TFPC features with different transformation approaches applied under (a) clean-speech training and (b) multi-condition training.

better then MFCC under matched environments [5, 6], for the current experimental environments are mismatched. However, when further processed with CMS, both TFPC features and MFCC provided better performance under both clean-speech and multi-condition training conditions, and the improvements for TFPC features were much more significant than those for MFCC in both cases. As a result, the accuracy gaps were narrowed, although MFCC are still slightly better than TFPC features. When we further replace the CMS with CN, a stronger feature transformation which not only subtracts the utterance-wise mean from a specific dimension of feature coefficients as CMS does but also divides them by corresponding standard deviation, the first important phenomenon worth noting is that now TFPC features surpass MFCC for both clean-speech training and multi-condition training. In addition, TFPC features achieve quite significant performance improvements for both clean-speech training and multi-condition training, as compared to the case when CMS is applied, while for MFCC a moderate increase can be found in accuracy for clean-speech training, but a slight degradation can also be observed for multi-condition training. These results indicated that stronger feature transformations may help TFPC features give better improvements than MFCC. Finally, when CMS is substituted by HEQ, the rightmost two bars in Figure 4 (a) and (b) show that further improvements for both MFCC and TFPC features can be easily observed in both clean-speech and multi-condition training. It should be noted that with HEQ the averaged performances of MFCC and TFPC features are quite similar under both training conditions. Though the accuracy of MFCC is slightly better under clean-speech training condition while TFPC features is slightly better under multi-condition training, the differences are not very significant.

Word Accuracy		MFCC+CN	TFPC+CN	MFCC+HEQ	TFPC+HEQ
Clean-speech training	Set A	70.28	76.26	80.38	78.67
	Set B	70.78	77.85	81.43	80.11
	Set C	66.37	77.29	80.81	78.77
	average	69.14	77.13	80.87	79.18
multi condition training	Set A	89.67	89.70	90.20	90.85
	Set B	87.95	89.54	89.62	90.59
	Set C	86.10	89.61	89.47	90.82
	average	87.81	89.62	89.76	90.75

Table 1: Word accuracy for MFCC and TFPC features with CN or HEQ applied under clean-speech and multi-condition training for different testing sets averaged over all SNR values.

Because the rightmost bars in Figure 4 (a) and (b) are the average of many different cases, we need to compare the results more carefully considering different cases. Here we first compare MFCC and TFPC features both with HEQ but for different testing sets A, B, and C, averaged over all SNR values. The results are listed in Table 1. From Table 1, it can be found that the general trend for the averaged accuracy in Figure 4 actually holds for individual testing sets A, B, and C. The TFPC features are better than MFCC when CN is applied, but becomes slightly worse if HEQ is applied under clean-speech training condition. However, these numbers in Table 1 are the averages over different SNR values. The situation is actually different for different SNR conditions. The word accuracies for MFCC and TFPC features with CN or HEQ applied, at different SNR's but averaged over all testing sets A, B, and C under clean-speech training and multi-condition training are plotted in Figure 5(a) and (b), respectively. Now from Figure 5 we clearly see why TFPC features with HEQ becomes worse than MFCC with HEQ with clean-speech training on average. In Figure 5(a) for clean-speech training, we can observe that as long as the SNR is 10 dB or higher, TFPC features with either CN or HEQ are always better than MFCC with HEQ. However, when SNR is only 5 or 0 dB, MFCC with HEQ become the best. These two cases of 5 and 0 dB are responsible for the differences in the rightmost bars of Figure 4 and the upper right quarter of Table 1. For the multi-condition training, in Figure 5(b), however, it is clear that TFPC features with HEQ are always the best for all SNR values. With all the observations mentioned above, it may be concluded that unless training and testing conditions are very seriously mismatched (e.g. SNR of 5 or 0 dB for the clean-speech training condition here), TFPC features integrated with HEQ transformation can offer very good robustness under adverse environments.

All the above results always include performance for different types of noise averaged together. Below in Figure 6 (a) and (b) we thus pick up two types of noise, exhibition noise and train station noise respectively, and show the results for each type of noise alone. The results in Figure 6 are quite consistent with the previous overall average, i.e., TFPC features with HEQ are always better than TFPC features with CN. Note that the

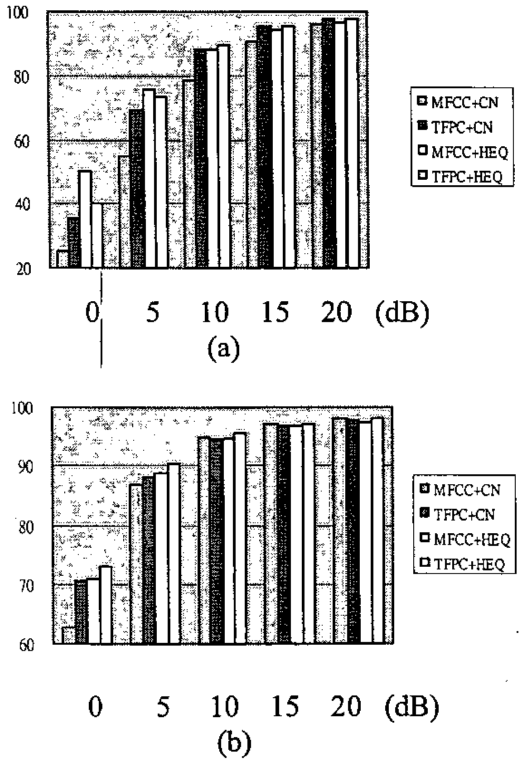


Figure 5: Average word accuracy for MFCC and TFPC features with CN or HEQ applied under (a) clean-speech training and (b) multi-condition training at different SNR values averaged over testing sets A, B, and C.

two types of noise chosen here are quite typical. The exhibition noise is very stationary and is within the multi-condition training data, while the train station noise is quite non-stationary and not included in the multi-condition training data. Also, TFPC features with CN or HEQ have comparable performance when the noise level is low, but the superiority of HEQ over CN becomes apparent when the noise level is higher.

4.3. Further improvements for TFPC integrated with HEQ by multi-eigenvector temporal filtering

As observed previously, TFPC features with HEQ turns out to be quite robust with respect to adverse environments, except it may become slightly weaker for seriously mismatched condition. Here we try to develop some extra approach to handle this limitation.

The data-driven temporal filtering approaches have been proved to be able to enhance the performance of CN-processed features under noisy conditions [7, 8], and a new multi-eigenvector temporal filtering approach was recently proposed and shown to provide very significant improvements in seriously mismatched environments [9]. In this approach, the temporal filters are obtained by the linear combination of m eigenvectors corresponding to the m largest eigenvalues, weighted by the eigenvalues, where the eigenvalues and eigenvectors are obtained with the covariance matrix for the segments of the time trajectories of the features. Compared to the conventional PCA temporal filtering which includes only the projection of the time trajectories onto the first principal component, the new multi-eigenvector filter used here tries to include the projections of the time trajectories onto the m major principal components, which also convey important information.

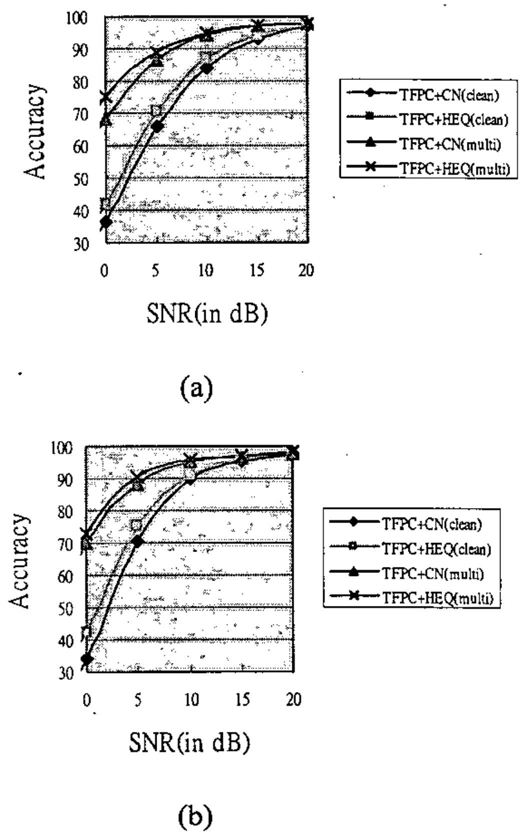


Figure 6: Word accuracy for TFPC features with CN and HEQ with clean-speech and multi-condition training for (a) exhibition noise and (b) train station noise.

In the preliminary experiments, this multi-eigenvector temporal filtering approach was applied to the HEQ-processed TFPC features, in which m=3 and the filter length was taken to be 8. The accuracies of testing sets A, B, and C averaged over all noise types and all SNR values for the clean-speech training condition are listed in table 2. Also, the accuracies at different SNR values averaged over the three testing sets A, B, and C, including those by TFPC alone, with CMS, CN, and HEQ, and finally enhanced by the multi-eigenvector approach for clean-speech training condition are plotted in Figure 7. The improvements obtainable with the multi-eigenvector temporal filtering as well as contributions of the various transformations to TFPC are all clear from Table 2 or Figure 7.

5. CONCLUSIONS

Approaches	Set A	Set B	Set C	Average
TFPC+HEQ	78.67	80.11	78.77	79.18
TFPC+HEQ+m-eign	81.35	81.92	81.79	81.69

Table 2: Word accuracy for multi-eigenvector temporal filtering applied to TFPC features with HEQ for clean-speech training.

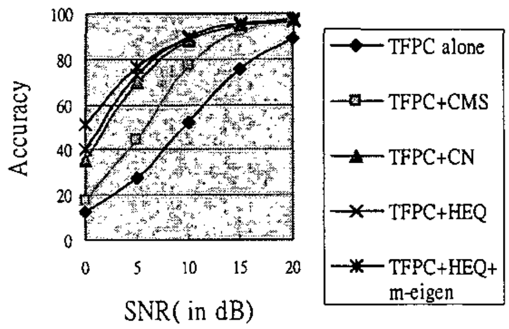


Figure 7: Word accuracy with different techniques applied to TFPC features under clean-speech training.

In this paper, we propose a new feature set which integrates the concept of TFPC and HEQ. Very promising experimental results were obtained with AURORA2 database, and even better word accuracy was achieved when the new features are combined with effective temporal filtering.

6. REFERENCES

[1] L.-s. Lee and Y. Lee, "Voice Access of Global Information for Broad-band Wireless: Technologies of Today and Challenges of Tomorrow", Proceedings of the IEEE, Jan 2001.

[2] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral Channel Normalization Techniques for HMM-based Speaker Verification", ICSLP, 1992.

[3] O. Viikki and K. Laurila, "Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature vector Normalization", in ESCA NATO Workshop Robust Speech Recognition Unknown Communication Channels, Pont-a-Mousson, France 1997.

[4] Á. de la Torre, J. C. Segura, C. Benítez, A. M. Peinado, and A. J. Rubio, "Non-linear Transformations of the Feature Space for Robust Speech Recognition", ICASSP, 2002.

[5] I. Magrin-Chagnolleau, and G. Durou, "Application of Time-Frequency Principal Component Analysis to Text-Independent Speaker Identification", IEEE, Trans. Speech and Audio Processing, Sep. 2002.

[6] P. Somervuo, "Experiments with Linear and Nonlinear Feature Transformations in HMM Based Phone Recognition", ICASSP 2003.

[7] S. van Vuuren and H. Hermansky, "Data-driven Design of RASTA-like Filters", Europe Speech 97.

[8] J.-h. Hung and L.-s. Lee, "Data-Driven Temporal Filters for Robust Features in Speech Recognition Obtained via Different Optimization Criteria Evaluated on Aurora2 Database", ICSLP, 2002.

[9] N.-c. Wang, J.-h. Hung and L.-s. Lee, "Data-Driven Temporal Filters Based on Multi-Eigenvectors for Robust Features in Speech Recognition", ICASSP, 2003.

[10] S. S. Kajarekar, B. Yegnanarayana and H. Hermansky, "A Study of Two Dimensional Linear Discriminants for ASR", ICASSP, 2001.

[11]I. T. Jolliffe, "Principal Components Analysis", Berlin, Germany: Springer-Verlag, 1986.

[12]C. Nadeu, D. Macho, and J. Hernando, "Filtering the Time Sequence of Spectral Parameters for Speech Recognition", Speech Communication 22 (1997)

[13] V. Digalakis, L. Neumeyer, and M. Perakakis, "Quantization of Cepstrum Parameters for Speech Recognition over the World Wide Web", ICASSP, 1998