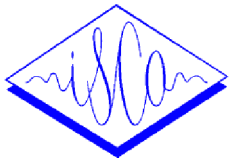

ITRW on Speech Recognition and
Intrinsic Variatioon (SRIV 2006)
Toulouse, France
May 20, 2006

Comparative study of different features on OLLO logatome recognition task

Vivek Tyagi, Mohammed Benzheghiba, Milos Cernak and Christian Wellekens,

Institute Eurecom, Sophia Antipolis, France
tyagi@eurecom.fr

Abstract

We compare the ASR performances of different features sets (MFCC, PLP, constant JRASTA PLP and variable scale piece-wise quasi-stationary analyzed MFCC features [1]) on the OLdenburg LOgatome speech corpus (OLLO)[2]. OLLO database is rich in various speech variabilities such as different speaking styles (slow, fast, statement, questioning, loud and soft) and with almost equal sampling of the male and female speakers. A HMM-GMM system has been trained on the no-accent part of the OLLO database that consists of roughly 13,500 utterances and then tested on the no-accent part of the test set that roughly consists of 13,800 utterances. Each of these utterances correspond to a logatome. We compare state-of the art fixed time scale (20ms long windows) features with the recently proposed variable scale quasi-stationary analyzed[1] MFCC features This technique results in a variable scale time spectral analysis, adaptively estimating the largest possible analysis window size such that the signal remains quasi-stationary, thus the best temporal/frequency resolution tradeoff. The speech recognition experiments on the OLLO database, show that the proposed variable-scale piecewise stationary spectral analysis based features indeed yield improved recognition accuracy in clean conditions, compared to MFCC, PLP and constant-JRASTA PLP features.

1. Introduction

In an information bearing signal such as speech or image, the information is propagated through the slow evolution of one quasi-stationary segment into another. For instance vowels slowly evolve to consonants and vice versa. The current ASR systems make a simplified assumption that all the stationary events are of uniform duration and the duration is typically assumed to be $20ms$. This poses a major limitation as certain sounds (events) such as vowels last for typically $(60ms - 80ms)$ while certain short-time-limited sounds such as plosive and stop last for $10ms$. The specific instants in a signal waveform when this stationarity switching happens, the rate at which this switching occurs and the duration of sustained stationary segments are all very important quantities which need to be detected and estimated to extract all the useful information from the speech signal

Most of the Automatic Speech Recognition (ASR) acoustic features, such as Mel-Frequency Cepstral Coefficient (MFCC) or Perceptual Linear Prediction (PLP), are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows of typically 20ms to 30ms of the speech signal [13]. Such analysis is based on the assumption that the speech signal can be assumed to be quasi-stationary over these segment durations. However, it is well known that the voiced speech sounds such as vowels are quasi-stationary for 40ms-80ms while, stops and plosive are time-limited by less than 20ms [13]. Therefore, it implies that the spectral analysis based on a fixed size window of 20ms-30ms has some limitations, including:

- The frequency resolution obtained for quasi-stationary segments (QSS) longer than 20ms is quite low compared to what could be obtained using larger analysis windows.
- In certain cases, the analysis window can span the transition between two QSSs, thus blurring the spectral properties of the QSSs, as well as of the transitions. Indeed, in theory, Power Spectral Density (PSD) cannot even be defined for such non stationary segments [9]. Furthermore, on a more practical note, the feature vectors extracted from such transition segments do not belong to a single unique (stationary) class and may lead to poor discrimination in a pattern recognition problem.

In this work, we make the usual assumption that the piecewise quasi-stationary segments (QSS) of the speech signal can be modeled by a Gaussian AR process of a fixed order p as in [6, 10, 11]. We then formulate the problem of detecting QSSs as a Maximum Likelihood (ML) detection problem, defining a QSSs as the longest segment that has most probably been generated by the same AR process. As is well known, given a p^{th} order AR Gaussian QSS, the Minimum Mean Square Error (MMSE) linear prediction (LP) filter parameters $[a(1),\ a(2), ...\ a(p)]$ are the most "compact" representation of that QSS amongst all the p^{th} order all pole filters [9]. In other words, the normalized "coding error"[1]

[1]The power of the residual signal normalized by the number of sam-

is minimum amongst all the p^{th} order LP filters. When erroneously analyzing two distinct p^{th} order AR Gaussian QSSs in the same non-stationary analysis window, it can be shown that the "coding error" will then always be greater than the ones resulting of QSSs analyzed individually in stationary windows[12]. As further explained in the next sections, this forms the basis of our criteria to detect piecewise quasi-stationary segments. Once the "start" and the "end" points of a QSS are known, all the speech samples coming from this QSS are analyzed within that window, resulting in (variable-scale) acoustic vectors.

The main contribution of the present paper is to demonstrate that the variable-scale QSS spectral analysis technique can possibly improve the ASR performance as compared to the fixed scale spectrum analysis. We do a comparative study of the proposed variable-scale spectrum based features and other state-of-the art features such as MFCC[3], PLP[5], C-JRASTA PLP[4]. In the following sections we will describe the PLP[5], JRASTA-PLP[4] and the variable scale piece-wise quasi-stationary analyzed[1] MFCC features followed by logatome recognition experiments on the OLLO[2] database using these features.

2. Perceptual Linear Prediction (PLP) features

Over the past few decades, many variants of filter banks, LPC, and cepstral vectors haven been used for speech recognition. More recently, the majority of the systems have converged to the use of a cepstral vector derived from a filter bank that has been designed according to some model of the auditory system. In the following we will briefly describe some of the auditory inspired steps involved in the PLP[5] feature computation.

- Compute a power spectral estimate for the analysis window;typically this is done by windowing the analysis segment, computing the FFT, and computing its squared magnitude to get the power spectrum.

- Integrate the power spectrum within overlapping critical band filter responses. There are number of forms used for these filters, but all of them are based on a frequency scale that is roughly linear below 1Khz and logarithmic above this point. The Mel scale is based on the pitch perception and is used in the filter banks for the MFCC approach. Since it is based on human experimental data, there are number of approximations and models that have been used. In the mel case, the integration step is done with a triangular window applied to the log of the power spectrum. For the case of PLP trapezoidally shaped filters are applied at roughly 1-Bark intervals. The trapezoidally shaped windows are an approximation to the power spectrum of the critical band masking curve from Fletcher. In both cases, the net effect is to reduce the frequency sensitivity over the original spectral estimate, particularly at higher frequencies. The higher frequencies are also somewhat emphasized given the wider filter bandwidths.

- Pre-emphasize the spectrum to approximate the unequal sensitivity of human hearing at different frequencies. In most mel-cepstral analysis, this is actually done before the original spectral analysis, and an important side effect is to eliminate the effects of the DC offsets in the speech signal.

- Compress the spectral amplitudes. Typically the log is applied after the integration. In PLP, the cube root is taken rather than the log which is an approximation to the power-law relationship between intensity and loudness.

- Perform an inverse DFT. It is a critical step for both MFCC and PLP. In the former case, it is the step that yields the cepstral coefficients. For PLP, since the log has not been computed, the results are more like autocorrelation like features(though they are still from a compressed spectrum.)

- Perform spectral smoothing. Although the critical band spectrum suppresses some detail, another level of integration has been shown to be useful for reducing the effects of non-linguistic sources of variance in the speech signal. In MFCC this step is accomplished by cepstral truncation; typically the lower 13 cepstral components are retained from 24 filter bank energies. Thus the resulting representation corresponds to a smoothed spectrum. In the case of PLP, an auto-regressive (derived by the solution of linear equations constructed from the autocorrelation of the previous step) is used to smooth the compressed critical band spectrum; as with conventional LPC, the resulting smoothed spectrum is a better fit to the spectral peaks than the valleys.

- use orthogonal representation. For MFCC no further step is necessary to get orthogonal features-the elements of the truncated cepstral vectors have this property. For PLP, the autoregressive coefficients are converted to cepstral coefficients.

3. Rasta processing

RASTA processing[4] tries to make speech analysis less sensitive to the slowly changing or steady-state factors in a speech signal. It replaces the conventional critical

ples in the window

band short term spectrum estimate of the PLP analysis with a spectral estimate in which each frequency channel is band-pass filtered by a filter with a sharp zero at zero frequency. Since any constant or slowly varying component in each frequency channel is suppressed by this operation, the new spectral estimate is less sensitive to the slow variations in the short-term spectrum. The bandpass IIR filter used has a low-cutoff frequency at 0.26 Hz. The filter slope declines 6db/octave from 12.8 Hz with sharp zeros at 28.9 and 50 Hz.

4. Detecting stationarity change over point in an auto-regressive signal

Consider an instance of a p^{th} order AR Gaussian process, $\mathbf{x[n]}, n \in [1, N]$ whose generative LP filter parameters can either be $\mathbf{A}_0 = [1, a_0(1), a_0(2)....a_0(p)]$ or can change from $\mathbf{A}_1 = [1, a_1(1), a_1(2)....a_1(p)]$ to $\mathbf{A}_2 = [1, a_2(1), a_2(2)....a_2(p)]$ at time n_1 where $n_1 \in [1, N]$. As usual, the excitation signal is assumed to be drawn from a white Gaussian process and its power can change from $\sigma = \sigma_1$ to $\sigma = \sigma_2$. The general form of the Power Spectral Density (PSD) of this signal is then known to be

$$P_{xx}(f) = \frac{\sigma^2}{|1 - \sum_{i=1}^{p} a(i) \exp(-j2\pi i f)|^2} \quad (1)$$

where $a(i)$s are the LPC parameters. The hypothesis test consists of:

- $\mathbf{H}_0$: No change in the PSD of the signal $x(n)$ over all $n \in [1, N]$, LP filter parameters are $\mathbf{A}_0$ and the excitation (residual) signal power is σ_0.
- $\mathbf{H}_1$: Change in the PSD of the signal $x(n)$ at n_1, where $n_1 \in [1, N]$, LP filter parameters change from $\mathbf{A}_1$ to $\mathbf{A}_2$ and the excitation(residual) signal power changes from σ_1 to σ_2.

Let, $\hat{\mathbf{A}}_0$ denote the maximum likelihood estimate (MLE) of the LP filter parameters and $\hat{\sigma}_0$ denote the MLE of the residual signal power under the hypothesis $\mathbf{H}_0$. The MLE estimate of the filter parameters is equal to their MMSE estimate due to the Gaussian distribution assumption [6] and, hence, can be computed using the Levinson Durbin algorithm [9] without significant computational cost.

Let $\mathbf{x}_1$ denote $[x(1), x(2), ...x(n_1)]$ and $\mathbf{x}_2$ denote $[x(n_1 + 1), ...x(N)]$. Under hypothesis $\mathbf{H}_1$, $(\hat{\mathbf{A}}_1, \hat{\sigma}_1)$ are the MLE of $(\mathbf{A}_1, \sigma_1)$ estimated on $\mathbf{x}_1$, and $(\hat{\mathbf{A}}_2, \hat{\sigma}_2)$ are the MLE of $(\mathbf{A}_2, \sigma_2)$ estimated on $\mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ have been assumed to be independent of each other. A Generalized Likelihood Ratio Test (GLRT) [12] would then pick hypothesis $\mathbf{H}_1$ if

$$\log L(\mathbf{x}) = \log\left(\frac{p(\mathbf{x}_1|\hat{\mathbf{A}}_1, \hat{\sigma}_1)p(\mathbf{x}_2|\hat{\mathbf{A}}_2, \hat{\sigma}_2)}{p(\mathbf{x}|\hat{\mathbf{A}}_0, \hat{\sigma}_0)}\right) > \gamma \quad (2)$$

where γ is a decision threshold that will have to be tuned on some development set. In [1], we have shown that (2) simplifies to the following,

$$\log L(\mathbf{x}) = \frac{1}{2} \log \left[\frac{\hat{\sigma}_0^N}{\hat{\sigma}_1^{n_1} \hat{\sigma}_2^{(N-n_1)}} \right] \quad (3)$$

In the present form, the GLRT $\log L(\mathbf{x})$ has now a natural interpretation. Indeed, if there is a transition point in the segment $\mathbf{x}$ then it has, in effect, $2p$ degrees of freedom. Under hypothesis $\mathbf{H}_0$, we encode $\mathbf{x}$ using only p degrees of freedom (LP parameters $\hat{\mathbf{A}}_0$) and, therefore, the coding (residual) error $\hat{\sigma}_0^2$ will be high. However, under hypothesis $\mathbf{H}_1$, we use $2p$ degrees of freedom (LP parameters $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$) to encode $\mathbf{x}$. Therefore, the coding (residual) errors $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be minimized to reach the lowest possible value.[2] This will result in $L(\mathbf{x}) > 1$. On the other hand, if there is no AR switching point in the segment $\mathbf{x}$ then it can be shown that, for large n_1 and N, the coding errors are all equal ($\hat{\sigma}_0^2 = \hat{\sigma}_1^2 = \hat{\sigma}_2^2$). This will result in $L(\mathbf{x}) \simeq 1$.

Brandt[10] had derived (2) and it was later on used by Obrect[11] for segmenting a speech signal into phonemes. Obrect[11] had reported that on an average their algorithm segments a phoneme into 2.2 segments per phoneme. However they do not show any relation of the ML detection to the variable scale quasi-stationary spectral analysis of speech signals and its extension for improved speech recognition as has been done in this work.

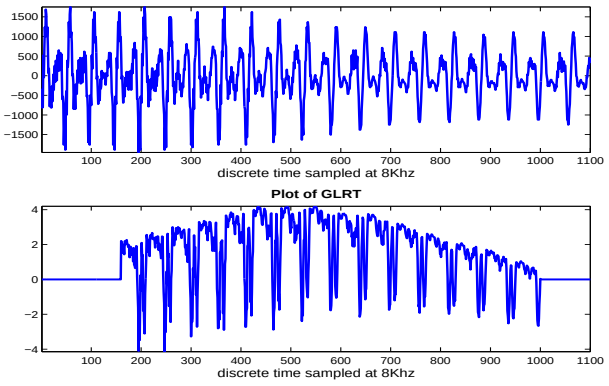

Figure 1: *Typical plot of the Generalized log likelihood ratio test (GLRT) for a speech segment. The sharp downward spikes in the GLRT are due to the presence of a glottal pulse at the beginning of the right analysis window ($\mathbf{x}_2$). The GLRT peaks around the sample 500 which marks as a strong AR model switching point*

An example is illustrated in Figure 1. The top pane shows a segment of a voiced speech signal. In the bottom

[2] When $\hat{\mathbf{A}}^1$ and $\hat{\mathbf{A}}^2$ are estimated, strictly based on the samples from the corresponding quasi-stationary segments.

figure, we plot the GLRT as the function of the hypothesized change over point n. Whenever, the right window i.e the segment $\mathbf{x}_2$ spans the glottal pulse in the beginning of the window, the GLRT exhibits strong downward spikes which is due to the fact that the LP filter cannot predict large samples in the beginning of the window. However, these downward spikes do not influence our decision significantly as we are interested in large positive value of the GLRT to detect a model change over point. The minimum sizes of the left and the right windows are 160 and 100 samples respectively. This explains the zero value of the GLRT at the beginning and the end of the whole test segment. The GLRT peaks around sample 500 which marks a strong AR model switching point.

5. Relation of GLRT to Spectral Matching

As is well known the LP error measure possesses the spectral matching property [7]. Specifically, given a speech segment $\mathbf{x}$, let its power spectrum (periodogram) be denoted by $\mathbf{X}(e^{j\omega})$. Let the all pole model spectrum of the segment $\mathbf{x}$ be denoted as $\hat{\mathbf{X}}_0(e^{j\omega})$. Then it can be shown that the MMSE error σ_0^2 of the LP filter estimated over the entire segment $\mathbf{x}$ is given by [7]

$$\sigma_0^2 = \int_{-\pi}^{\pi} \frac{\mathbf{X}(e^{j\omega})}{\hat{\mathbf{X}}_0(e^{j\omega})} d\omega \text{ where,} \quad (4)$$

$$\hat{\mathbf{X}}_0(e^{j\omega}) = \frac{1}{|1 - \sum_{i=1}^{p} a_0(i) \exp(-j2\pi i f)|^2} \quad (5)$$

Therefore minimizing the residual error σ_0^2 is equivalent to the minimization of the integrated ratio of the signal power spectrum $\mathbf{X}(e^{j\omega})$ to its approximation $\hat{\mathbf{X}}_0(e^{j\omega})$ [7]. Substituting (4) in (3) we obtain,

$$\log L(\mathbf{x}) = \quad (6)$$
$$\frac{1}{2} \log \frac{\left(\int_{-\pi}^{\pi} \frac{\mathbf{X}(e^{j\omega})}{\hat{\mathbf{X}}_0(e^{j\omega})} d\omega \right)^N}{\left(\int_{-\pi}^{\pi} \frac{\mathbf{X}_1(e^{j\omega})}{\hat{\mathbf{X}}_1(e^{j\omega})} d\omega \right)^{n_1} \left(\int_{-\pi}^{\pi} \frac{\mathbf{X}_2(e^{j\omega})}{\hat{\mathbf{X}}_2(e^{j\omega})} d\omega \right)^{N-n_1}}$$

where, $\mathbf{X}(e^{j\omega})$, $\mathbf{X}_1(e^{j\omega})$ and $\mathbf{X}_2(e^{j\omega})$ are the power spectra of the segments $\mathbf{x}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. Similarly $\hat{\mathbf{X}}_0(e^{j\omega})$, $\hat{\mathbf{X}}_1(e^{j\omega})$ and $\hat{\mathbf{X}}_2(e^{j\omega})$ are the MMSE p^{th} order all-pole model spectra estimated over the segments $\mathbf{x}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. Therefore, $\hat{\mathbf{X}}_0(e^{j\omega})$, $\hat{\mathbf{X}}_1(e^{j\omega})$ and $\hat{\mathbf{X}}_2(e^{j\omega})$ are the best spectral matches to their corresponding power spectra. One way of interpreting (6) is that it is a measure of the relative goodness between the best spectral match achieved by modeling $\mathbf{x}$ as a single QSS and the best spectral matches obtained by assuming $\mathbf{x}$ to consist of two distinct QSS, namely $\mathbf{x}_1$ and $\mathbf{x}_2$. This is further explained as follows. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are indeed two distinct QSS, then $\mathbf{X}_1(e^{j\omega})$ and $\mathbf{X}_2(e^{j\omega})$ will be quite different and $\mathbf{X}(e^{j\omega})$ will be a gross average of these two spectra. In other words, the frequency support of $\mathbf{X}(e^{j\omega})$ will be a union of those of the $\mathbf{X}_1(e^{j\omega})$ and $\mathbf{X}_2(e^{j\omega})$. $\hat{\mathbf{X}}_1(e^{j\omega})$ and $\hat{\mathbf{X}}_2(e^{j\omega})$, having p poles each, will match their corresponding power spectra reasonably well, resulting in a lower value of the denominator in (6). However, $\hat{\mathbf{X}}_0(e^{j\omega})$ will be a relatively poorer spectral match to $\mathbf{X}(e^{j\omega})$ as it has only p poles to account for the wider frequency support. Therefore we incur a higher spectral mismatch by assuming $\mathbf{x}$ to be a single QSS when in fact it is composed of two distinct QSS $\mathbf{x}_1$ and $\mathbf{x}_2$. This results in the GLRT $\log L(\mathbf{x})$ taking up a high value. Whereas if $\mathbf{x}_1$ and $\mathbf{x}_2$ are the instances of the same quasi-stationary process, then so is $\mathbf{x}$. Therefore $\mathbf{X}_1(e^{j\omega})$, $\mathbf{X}_2(e^{j\omega})$ and $\mathbf{X}(e^{j\omega})$ are nearly the same with similar all-pole models, resulting in a value of the GLRT close to zero. The above discussion points out to the fact that the QSS analysis based on the proposed GLRT is constantly striving to achieve a better time varying spectral modeling of the underlying signal as compared to single fixed scale spectral analysis.

6. Experiments and Results

Table 1: *Logatome recognition rate over the entire variabilities.*

MFCC 20ms	72.50
PLP 20ms	73.04
Constant JRASTA PLP	69.72
Proposed Variable-scale QSS MFCC	75.34

We have used the GLRT L($\mathbf{x}$) in (3) to perform QSS spectral analysis of speech signals for ASR applications. We initialize the algorithm with a left window size `W_L` = 20ms and a right window size `W_R` = 10ms. We compute their corresponding MMSE residuals and the MMSE residual of the union of the two windows. Then, the GLRT is computed using (3) and is compared to the threshold. The parameter threshold $\gamma = 4.5$ was empirically tuned to achieve the best recognition accuracy. In general, the ASR results are slightly sensitive to the threshold, although not in a huge way. If the GLRT is greater than the threshold γ, `W_L` is considered the largest possible QSS and we obtain a spectral estimate using all the samples in `W_L`. Otherwise,`W_L` is incremented by `INCR`=0.625ms and the whole process is repeated until GLRT exceeds γ or `W_L` becomes equal to the maximum window size `WMAX`=$50ms$. The computation of a MFCC feature vector from a very small segment (such as 10ms) is inherently very noisy.[3] Therefore, the minimum duration of a QSS as detected by the algorithm was constrained to be $20ms$. Throughout the experiments, a fixed LP order $p = 10$ was used. To avoid fluctuating Nyquist frequency of the cepstral modulation spectrum[8], a fixed shift size of $10ms$ was used in the algorithm. Comparative study of the different feature sets was performed over

[3] Due to very few samples involved in the Mel-filter integration.

Table 2: *Logatome recognition rate reported over each variabilities.*

Feature	Fast	Slow	Loud	Soft	Questioning	Normal
MFCC 20ms	68.14	73.02	73.35	63.75	78.49	78.21
PLP 20ms	67.56	76.80	72.15	65.85	77.37	78.48
Constant JRASTA PLP	57.09	75.47	68.95	64.24	77.55	75.00
Proposed Variable-scale QSS MFCC	70.23	76.54	75.93	67.23	80.49	81.60

the OLLO database. In these experiments we decided to recognize the entire logatome. The lexicon size is 150 that corresponds to 150 logatomes.

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK on the NO-accent part of the OLLO[2] training set that roughly consist of 13,500 utterances. Three state left to right HMM models were trained for each of the 26 phonemes in the OLLO[2] database including silence as well. The lexicon consists of 150 logatomes which are either CVCs or VCVs. The experiments reported in this paper are for the entire logatome recognition on the No-accent test part of the OLLO database that consists of roughly 13,800 utterances. Each of these utterances correspond to an instance of a logatome. Gaussian mixture models(GMMS) with 25 Gaussian per state and diagonal covariance matrices were used to model the emission probability densities of the feature vectors. The number of parameters used in the HMM-GMM system is the same for all the features reported. The logatome recognition results for various features are given in Tables 1, 2.

As can be noted from the Table 2, C-JRASTA PLP has drastically poor performance on fast speech as compared to the rest of the features that may be due to the fact that RASTA processing involves band-pass filtering of filter bank modulation energies in the range [2,8]Hz. This may also imply that the range of the modulation frequencies that correspond to speech is probably a function of the speaking rate. We note that the variable scale QSS analyzed MFCC have better recognition accuracy as compared to the fixed scale MFCC for all the six variabilities. However, we note that PLP feature performs significantly better than the MFCC feature on the slow speech.(76.8 % of PLP Vs 73.02 % of MFCC Vs 75.47 % of C-JRASTA PLP). However the variable scale QSS analyzed MFCC compensates for this deficiency of the fixed scale MFCC by reaching an accuracy of 76.54 % on slow speech.

7. Conclusion

We have demonstrated that the variable-scale piecewise quasi-stationary spectral analysis of speech signal can possibly improve the state-of-the-art ASR. Such a technique can overcome the time-frequency resolution limitations of the fixed scale spectral analysis techniques. Comparisons were drawn with the other state-of-the art features based on the speech recognition accuracies.

8. Acknowledgment

This work has been supported by EC 6^{th} Framework project DIVINES under the contract number FP6-002034.

9. References

[1] Vivek Tyagi, Christian Wellekens and Herve Bourlard, "A Variable-Scale Piecewise Stationary Spectral Analysis Technique Applied to ASR," In Springer Lecture Notes in Computer Science LNCS-3869, Eds. Steve Renals and Samy Bengio, MLMI-2005, Edinburgh UK.

[2] T Wesker, B. Meyer, K. Wagener, J. Anemuller, A. Mertins, B. Kollmeier, "Oldenburg Logatome speech corpus (OLLO) for speech recognition experiments with humans and machines." In the Proceedings of ICSLP 2005, Lisbon, Portugal.

[3] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, " IEEE Trans. on ASSP, Vol. ASSP-28, No. 4, August 1980.

[4] H. Hermansky, N. Morgan, " Rasta Processing of Speech," IEEE Trans. on SAP, vol.2, no.4, October 1994.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech, " J. Acoust. Soc. Am., vol.87:4, April 1990.

[6] F. Itakura, " Minimum Prediction Residual Principle Applied to Speech Recognition, " IEEE Trans. on ASSP, Vol.23, no.1, February 1975.

[7] J. Makhoul, "Linear Prediction: A Tutorial Review, " In the Proc. of IEEE, vol.63, No.4, April 1975.

[8] V. Tyagi, I McCowan, H. Bourlard, H. Misra, " Mel-Cepstrum Modulation Spectrum (MCMS) features for Robust ASR, " In the Proc. of IEEE ASRU 2003, St. Thomas, Virgin Islands, USA.

[9] S. Haykin, Adaptive Filter Theory, Prentice-Hall Publishers, N.J., USA, 1993.

[10] A. V. Brandt, "Detecting and estimating the parameters jumps using ladder algorithms and likelihood ratio test," in Proc. of ICASSP, Boston, MA, 1983,pp. 1017-1020.

[11] R. A. Obrecht, "A new Statistical Approach for the Automatic Segmentation of Continuous Speech Signals, " IEEE Trans. on ASSP, vol.36, No.1, January 1988.

[12] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Prentice-Hall Publishers, N.J., USA, 1998.

[13] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, N.J., USA, 1993.