IDIAP RESEARCH REPORT

On Variable-Scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR

Vivek Tyagi [a,c] Hervé Bourlard [b,c]
Christian Wellekens [a,c]

IDIAP–RR 05-19

May 2005

submitted for publication

[a] Eurecom Institute, P.O Box 193, Sophia-Antipolis, France
[b] IDIAP Research Institute, P.O. Box 592, Martigny, Switzerland
[c] Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-19

On Variable-Scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR

Vivek Tyagi Hervé Bourlard Christian Wellekens

May 2005

submitted for publication

Abstract. It is often acknowledged that speech signals contain short-term and long-term temporal properties that are difficult to capture and model by using the usual fixed scale (typically 20ms) short time spectral analysis used in hidden Markov models (HMMs), based on piecewise stationarity and state conditional independence assumptions of acoustic vectors. For example, vowels are typically quasi-stationary over 40-80ms segments, while plosives typically require analysis below 20ms segments. Thus, fixed scale analysis is clearly sub-optimal for "optimal" time-frequency resolution and modeling of different stationary phones found in the speech signal. In the present paper, we investigate the potential advantages of using variable size analysis windows towards improving state-of-the-art speech recognition systems. Based on the usual assumption that the speech signal can be modeled by a varying autoregressive (AR) Gaussian process, we estimate the largest piecewise quasi-stationary speech segments, based on the likelihood that a segment was generated by the same AR process. This likelihood is estimated from the Linear Prediction (LP) residual error. Each of these quasi-stationary segments is then used as an analysis window from which spectral features are extracted. Such an approach thus results in 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. Speech recognition experiments on the OGI Numbers95 database show that the proposed multi-scale piecewise stationary spectral analysis based features indeed yield improved recognition accuracy in clean conditions, compared to features based on minimum cross entropy spectrum as well as those based on fixed scale spectral analysis.

1 Introduction

Most of the Automatic Speech Recognition (ASR) acoustic features, such as Mel-Frequency Cepstral Coefficient (MFCC)[16] or Perceptual Linear Prediction (PLP)[17], 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 [16, 15]. 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 [15]. 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 [2, 8, 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.[1] 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"[2] 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[14]. This is intuitively satisfying since, in the former case, we are trying to encode $'2p'$ free parameters (the LP filter coefficients of each of the QSS) using only p parameters (as the two distinct QSS are now analyzed within the same window). Therefore, higher coding error is expected in the former case as compared to the optimal case when each QSS is analyzed in a stationary window. 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.

Our algorithm is thus reminiscent of the likelihood ratio test based ML segmentation algorithm derived by Brandt [10] and later on used in [11]. In [11], the author has illustrated certain speech waveforms with segmentation boundaries overlaid. The validity of their algorithm is shown by a segmentation experiment, which on an average, segments phonemes into 2.2 segments. This result is quite useful as a pre-processor for the manual transcription of speech signals. However, the author in [11] did not discuss or extend the ML segmentation algorithm as a variable-scale quasi-stationary spectral analysis technique suitable for ASR, as done in the present work.

Before proceeding further, however, we feel necessary to briefly discuss certain inconsistencies between variable-scale spectral analysis and state-of-the-art Hidden Markov models ASR using Gaussian mixture models (HMM-GMM). HMM-GMM systems typically use spectral features based on a constant window size (typically $20ms$) and a constant shift size (typically $10ms$). The shift size determines the Nyquist frequency of the cepstral modulation spectrum [6], which is typically measured by

[1] Equivalent to the detection of the transition point between the two adjoining QSSs.

[2] The power of the residual signal normalized by the number of samples in the window

the delta features of the static MFCC or PLP features. In a variable-scale piecewise quasi-stationary analysis, the shift size should preferably be equal to the size of the detected QSS. Otherwise, if the shift size is x% of the duration of the QSS, then the next detected QSS will be the same but of duration $(100 - x)\%$ and the following one will be of duration $(100 - 2x)\%$ and so on until we have shifted past the entire duration of the QSS. This results in the undesirable effect that the same QSS gets analyzed by successively smaller windows, hence increasing the variance of the feature vector of this QSS. On the other hand, the use of a shift size equal to the variable window size will change the Nyquist frequency of the cepstral modulation spectrum [6]. Therefore, the modulation frequency pass-band of the delta filters [6] will vary from frame to frame and may suffer from aliasing for shift sizes in excess of $20ms$.

In [3], Atal has described a temporal decomposition technique to represent the continuous variation of the LPC parameters as a linearly weighted sum of a number of discrete elementary components. These elementary components are designed such that they have the minimum temporal spread (highly localized in time) resulting in superior coding efficiency. However, the relationship between the optimization criterion of "the minimum temporal spread" and the quasi-stationarity is not obvious. Therefore, the discrete elementary components are not necessarily quasi-stationary and vice-versa.

Coifman et al [5] have described a minimum entropy basis selection algorithm to achieve the minimum information cost of a signal relative to the designed orthonormal basis. In [7], Srinivasan et. al. have proposed a multi-scale QSS speech enhancement technique based on Coifman's technique [5]. In [8], Svendsen et al have proposed a ML segmentation algorithm using a single fixed window size for speech analysis, followed by a clustering of the frames which were spectrally similar for sub-word unit design. We emphasize here that this is different from the approach proposed here where we use variable size windows to achieve the objective of piecewise quasi-stationary spectral analysis. More recently, Achan et al [13] have proposed a segmental HMM for speech waveforms which identifies waveform samples at the boundaries between glottal pulse periods with applications in pitch estimation and time-scale modifications.

Our emphasis in this paper is on better spectral modeling of the speech signal rather than achieving better coding efficiency or reduced information cost. Nevertheless, we believe that these two objectives are somewhat fundamentally related. 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 identify the above mentioned problems and make certain engineering design choices to overcome these problems. Moreover, we show the relationship between the maximum likelihood QSS detection algorithm and the well known spectral matching property of the LP error measure [4]. Finally, we do a comparative study of the proposed variable-scale spectrum based features and the minimum cross-entropy time-frequency distributions developed by Loughlin et al [1].

In the sequel of this paper, Section 2 formulates the ML detection problem for identifying the transition points between QSS. In Section 3, we illustrate an analogy of the proposed technique with spectral matching property of the LP error measure. Finally, the experimental setup and results are described in Section 4.

2 ML Detection of the change-point in an AR Gaussian random process

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

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

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 [2] 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 ($\mathbf{A}_1$, σ_1) estimated on $\mathbf{x}_1$, and ($\hat{\mathbf{A}}_2$, $\hat{\sigma}_2$) are the MLE of ($\mathbf{A}_2$, σ_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) [14] would then pick hypothesis $\mathbf{H}_1$ if

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

where γ is a decision threshold that will have to be tuned on some development set. Given that the total number of samples in $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same as in $\mathbf{x}_0$, their likelihoods can be compared directly in (2). Under the hypothesis $\mathbf{H}_0$ the entire segment $\mathbf{x} = [x(1)...x(N)]$ is considered stationary and the MLE $\hat{\mathbf{A}}_0$ is computed via the Levinson-Durbin algorithm using all the samples in segment $\mathbf{x}$. It can be shown that the MLE $\hat{\sigma}_0$ is the power of the residual signal [2, 14]. Under $\mathbf{H}_1$, we assume that there are two distinct QSS, namely $\mathbf{x}_1$ and $\mathbf{x}_2$. The MLE $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ are computed via the Levinson-Durbin algorithm using samples from their corresponding QSS. MLE $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are computed as the power of the corresponding residual signals. In fact, $p(\mathbf{x}|\hat{\mathbf{A}}_0, \hat{\sigma}_0)$ is equal to the probability of residual signal reconstructed using filter parameters $\hat{\mathbf{A}}_0$, yielding:

$$p(\mathbf{x}|\hat{\mathbf{A}}_0, \hat{\sigma}_0) \quad = \quad \frac{1}{(2\pi\hat{\sigma}_0^2)^{N/2}} \exp\left[\frac{-1}{2\hat{\sigma}_0^2}\sum_{n=1}^{N}(e_0^2(n))\right] \tag{3}$$

where $e_0(n)$ is the residual error and

$$e_0(n) = x(n) - \sum_{i=1}^{p} a_0(i)x(n-i), \;\; n \in [1, N]$$

and

$$\hat{\sigma}_0^2 = \frac{1}{N}\sum_{n=1}^{N} e_0^2(n))$$

Similarly, $p(\mathbf{x}_1|\hat{\mathbf{A}}_1, \hat{\sigma}_1)$ and $p(\mathbf{x}_2|\hat{\mathbf{A}}_2, \hat{\sigma}_2)$ are the likelihoods of the residual signal vectors of the AR models $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, and have the same functional forms as above. Substituting these expressions into (2) yields

$$\log L(\mathbf{x}) = \frac{1}{2}\log\left[\frac{\hat{\sigma}_0^N}{\hat{\sigma}_1^{n_1}\hat{\sigma}_2^{(N-n_1)}}\right] \tag{4}$$

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.[3] 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 $L(\mathbf{x}) \simeq 1$.

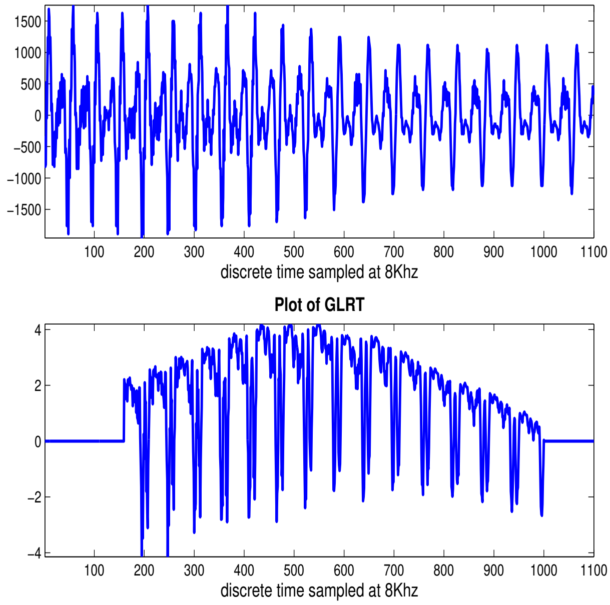

Figure 1: *Typical plot of the Generalized log likelihood ratio test (GLRT) for a speech segment. The sharp downward spikes in the GLRT are due to the presence of a glottal pulse at the beginning of the analysis window. The GLRT peaks around the sample 500 which marks as a strong AR model switching point*

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

3 Relation of GLRT to Spectral Matching

As is well known the LP error measure possesses the spectral matching property [4]. 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 [4]

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

$$\sigma_0^2 = \int_{-\pi}^{\pi} \frac{\mathbf{X}(e^{j\omega})}{\hat{\mathbf{X}}_0(e^{j\omega})} d\omega \quad \text{where,} \tag{5}$$

$$\hat{\mathbf{X}}_0(e^{j\omega}) = \frac{1}{|1 - \sum_{i=1}^{p} a_0(i) \exp(-j2\pi i f)\,|^2} \tag{6}$$

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 $\hat{\mathbf{X}}_0(e^{j\omega})$ [4]. Substituting (5) in (4) we obtain,

$$\log L(\mathbf{x}) = \frac{1}{2} \log \frac{\left(\int_{-\pi}^{\pi} \frac{\mathbf{X}(e^{j\omega})}{\hat{\mathbf{X}}_0(e^{j\omega})} d\omega \right)^N}{\left(\int_{-\pi}^{\pi} \frac{\mathbf{X}_1(e^{j\omega})}{\hat{\mathbf{X}}_1(e^{j\omega})} d\omega \right)^{n_1} \left(\int_{-\pi}^{\pi} \frac{\mathbf{X}_2(e^{j\omega})}{\hat{\mathbf{X}}_2(e^{j\omega})} d\omega \right)^{N-n_1}} \tag{7}$$

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 (7) 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 (7). 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.

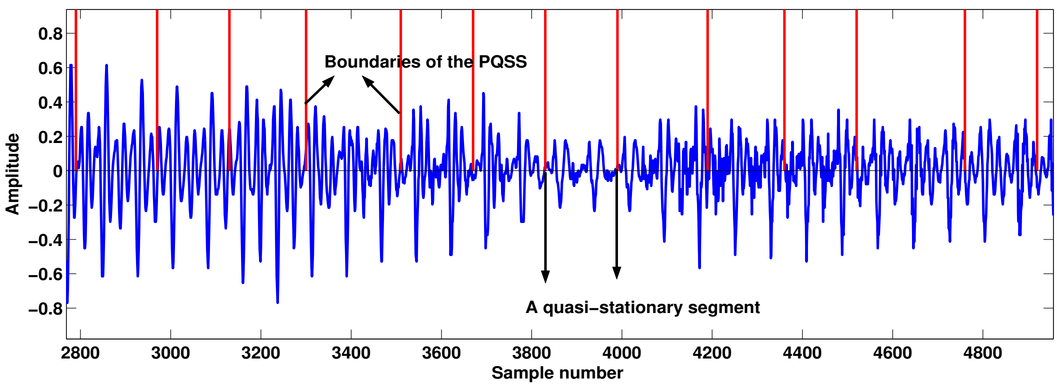

Figure 2: Quasi stationary segments (QSS) of a speech signal as detected by the algorithm with $\gamma = 3.5$ and LP order $p = 14$.

4 Experiments and Results

We have used the GLRT $L(\mathbf{x})$ in (4) 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` = 12.5ms. We compute their corresponding MMSE residuals and the MMSE residual of the union of the two windows. Then, the GLRT is computed using (4) and is compared to the threshold. The choice of the threshold $\gamma = 3.5$ was obtained by a visual inspection of the quasi-stationarity of the segmented speech signal as returned by the algorithm. In figure (2), we illustrate the boundaries of the QSS as detected by the algorithm with $\gamma = 3.5$. Realizing that the resulting segmentation corresponded to reasonably quasi-stationary segments, we adopted the threshold value $\gamma = 3.5$ for all the experiments reported in this paper. 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`=1.25ms and the whole process is repeated until GLRT exceeds γ or `W_L` becomes equal to the maximum window size `WMAX`=$60ms$. The computation of a MFCC feature vector from a very small segment (such as 10ms) is inherently very noisy.[4] Therefore, the minimum duration of a QSS as detected by the algorithm was constrained to be $20ms$. Throughout the experiments, a fixed LP order $p = 14$ was used.

The likelihood ratio test is quite widely used for speaker segmentation [12] where the average length of a single speaker segment may last from 1sec to several seconds. This provides a relatively large amount of samples to estimate the parameters of the probability density functions as compared to the present problem where we have to detect stationarity change over point within 20ms to 60ms. As a result, the GLRT in (4) is quite noisy and a criterion such as local maxima of the GLRT cannot be used. However, when the model change occurs over longer time scales (e.g. speaker change detection where the smallest segment is of the order of a second), Ajmera et al. [12] have successfully used a local maxima of the GLRT as a speaker change point detector, thereby, avoiding the use of a threshold.

To avoid fluctuating Nyquist frequency of the cepstral modulation spectrum[6], a fixed shift size of $12.5ms$ was used in the algorithm. As explained in the Section (1), this sometime resulted in the undesirable effect that the same QSS gets analyzed by progressively smaller windows. To alleviate this problem, the zeroth cepstral coefficient $c(0)$, which is a non-linear function of the windowed signal energy and, hence, of the window size, was normalized such that its dependence on the window size is mimimized.

In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the OGI Numbers corpus [19]. This database contains spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words. Figure (3) illustrates the distribution of the QSSs as detected by the proposed algorithm. Nearly 47% segments were analyzed with the smallest window size of $20ms$ and they mostly corresponded to short-time limited segments. However, voiced segments and long silences were mostly analyzed by using longer windows in the range $30ms - 60ms$. The short peak at 60ms is due to the accumulated value over all the segments that should have been longer than 60ms but were constrained by our choice of the largest window size.

Throughout the experiments, MFCC coefficients and their temporal derivatives were used as speech features. However, five feature sets were compared:

1. [39 dim. MFCC:] computed over a fixed window of length 20ms.
2. [39 dim. MFCC:] computed over a window of length 50ms.
3. [78 dim. Concatenated MFCC:] a concatenation of the above two feature vectors.
4. [Minimum cross entropy,39 dim MFCC:] MFCC computed from the geometric mean of the spectra computed from 20ms, 30ms, 40ms and 50ms long windows.

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

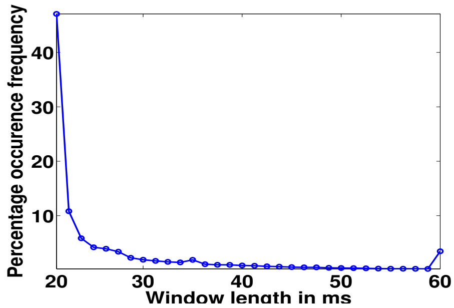

Figure 3: *Distribution of the QSS window sizes detected and then used in the training set*

5. [Variable-scale QSS MFCC+Deltas:] For a given frame, the window size is dynamically chosen using the proposed algorithm ensuring that the windowed segment is quasi-stationary.

In [1], Loughlin et al proposed using a geometric mean of multiple spectrograms of different window sizes to overcome the time-frequency limitation of any single spectrogram. They showed that combining the information content from multiple spectrograms in form of their geometric mean, is optimal for minimizing the cross entropy between the multiple spectra. We have followed their approach to derive MFCC features from the geometric mean of the multiple power spectra computed over varying window sizes, specifically 20ms, 30ms, 40ms and 50ms.

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK [18] on the clean training set from the original Numbers corpus. The speech recognition results in clean conditions for various spectral analysis techniques are given in table 1. The fixed scale MFCC features using 20ms and 50ms long analysis windows have 5.8% and 5.9% word error rate (WER) respectively. The concatenation of MFCC feature vectors derived from 20ms and 50ms long windows has a 5.7% WER and it has twice the number of HMM-GMM parameters as compared to the rest of the systems[5]. The slight improvement in this case may be due to the multiple scale information present in this feature, albeit in an ad-hoc way. The minimum cross-entropy MFCC features which were derived from the geometric mean of the power spectra computed over 20ms, 30ms, 40ms and 50ms long analysis windows, have a WER of 5.7%. The proposed variable-scale system which adaptively chooses a window size in the range [20ms, 60ms], followed by the usual MFCC computation, has a 5.0% WER. This corresponds to a relative improvement of more than 10% over the rest of the techniques

Table 1: *Word error rate in clean conditions*

MFCC 20ms	5.8
MFCC 50ms	5.9
Concat. MFCC (20ms, 50ms)	5.7
Min. Cross entropy based MFCC	5.7
Proposed Multi-scale QSS MFCC	5.0

5 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

[5]Due to twice the feature dimension as compared to the rest of the systems

other competing multi-scale techniques such as the minimum cross-entropy spectrum. The proposed technique led to the minimum WER as compared to the rest of the techniques.

References

[1] P. Loughlin, J. Pitton and B. Hannaford, "Approximating Time-Frequency Density Functions via Optimal Combinations of Spectrograms," IEEE Signal Processing Letters, vol.1, No.12, December 1994.

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

[3] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition, " In the Proc. of IEEE ICASSP, Boston, USA, 1983.

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

[5] R. R. Coifman and M. V. Wickerhauser, " Entropy based algorithms for best basis selection, " IEEE Trans. on Information Theory, Vol.38, Issue:2, March 1992.

[6] 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.

[7] S. Srinivasan and W. B. Kleijn, "Speech Enhancement Using Adaptive time-domain Segmentation, " In the Proc. of ICSLP 2004, Jeju, S. Korea.

[8] T. Svendsen, K. K. Paliwal, E. Harborg, P. O. Husoy, "An improved sub-word based speech recognizer," Proc. of IEEE ICASSP, 1989.

[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] J. Ajmera, I. McCowan and H. Boulard, "Robust Speaker Change Detection," IEEE Signal Processing Letters, vol.11, No. 8, August 2004.

[13] K. Achan, S. Roweis, A. Hertzmann and B. Frey, "A Segmental HMM for Speech Waveforms, " UTML Techical Report 2004-001, Dept. of Computer Science, Univ. of Toronto, May 2004.

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

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

[16] 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.

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

[18] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book, Cambridge University, 1995.

[19] R. A. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at CSLU," Proc. of ICSLP, Yokohama, Japan, 1994.