Institut Eurécom
Department of Multimedia Communications
2229, route des Crétes
B.P. 193
06904 Sophia-Antipolis
FRANCE

Research Report RR-04-118

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

Date 20th September 2004

Vivek Tyagi, Christian Wellekens

Tel : (+33) 4 93 00 26 26
Fax : (+33) 4 93 00 26 27
Email : {Vivek.Tyagi , Christian.Wellekens}@eurecom.fr

[1]Institut Eurécom's research is partially supported by its industrial members: Bouygues Téécom,Fondation d'entreprise Groupe Cegetel, Fondation Hasler, France Télécom, Hitachi, ST Microelectronics, Swisscom, Texas Instruments, Thales

Abstract

A £xed scale (typically 25ms) short time spectral analysis of speech signals, which are inherently multi-scale in nature [7] (typically vowels last for 40-80ms while stops last for 10-20ms), is clearly sub-optimal for time-frequency resolution. In this work, we detect piecewise quasi-stationary speech segments based on the likelihood of that segment which in turn is estimated from the linear prediction (LP) residual error. A window size equal in length to that of the detected quasi-stationary segment is used to obtain its spectral estimate. Such an approach adaptively chooses the largest possible window size such that the signal remains quasi-stationary within this window and excludes the adjoining quasi-stationary segments from this window. In experiments, it is shown that the proposed multi-scale piecewise stationary spectral analysis based features improve recognition accuracy in clean conditions when compared directly to features based on £xed scale spectral analysis.

1 Introduction

Speech signals as many other signals are inherently multi-scale in nature, owing to contributions from events occurring with different localizations in time and frequency. Therefore, signal analysis and modeling methods that represent the measured signal at multiple scales are better suited for extracting information from signal than methods that represent it at a single £xed scale.

Most of the front-ends (such as MFCC or PLP) used in current automatic speech recognition systems (ASR), employ a smoothed spectral envelope estimated over 20ms to 30ms of speech signal[10, 7]. This is based on the long-standing assumption that the speech signal can be assumed to be quasi-stationary for these 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 20ms [7]. Therefore, it implies that the spectral analysis based on a window of single £xed size (20ms-30ms) has the following serious limitations:

- The frequency resolution obtained for speech segments which are quasi-stationary for durations much longer than 20ms, is quite low as compared to what one can obtain using longer windows.
- In certain cases, more than one quasi-stationary segment (QSS) might be erroneously analyzed in the same analysis window (for instance, around the transition points between two QSSs). Power spectral density (PSD) cannot even be de£ned for such non stationary segments [1]. On a more practical note, the feature vectors extracted from such non stationary segments do not belong to a single unique class and may lead to poor discrimination in a pattern recognition problem.

In this work, we make the assumption that the piecewise quasi-stationary segments (PQSS) of the speech signal can be modeled by a Gaussian AR process of a £xed order $'p'$ as in [2]. We formulate the problem at hand as a ML detection of model change over point.[1] As is well known, given a p^{th} order AR Gaussian PQSS, the minimum mean square error (MMSE) linear prediction (LP) £lter parameters $[a(1),\ a(2), ...\ a(p)]$ are the most "compact" representation of that PQSS amongst all the p^{th} order all pole £lters [1]. In other words, the normalized[2] "coding error" is minimum amongst all the p^{th} order LP £lters. Now, consider a case when we erroneously analyze two distinct p^{th} order AR Gaussian PQSSs in the same non-stationary analysis window. Then, it can be shown that the "coding error" in this case is greater than the ones obtained when the two PQSSs are analyzed individually in stationary windows[6]. This is intuitively satisfying as in the former case we are trying to encode $'2p'$ free parameters (the LP £lter coef£cients of each of the PQSS) using only p parameters (as the two distinct PQSS are now analyzed

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

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

within the same window). Therefore higher coding error is expected in the former case as compared to the optimal case when each PQSS is analyzed in a stationary window. As it will be further explained in the later sections, this forms the basis of our criteria to detect piecewise quasi-stationary segments. Once the "start" and the "end" point of a PQSS are know, all the speech samples coming from this PQSS are analyzed within the same window. This can be seen as locking the windows to the PQSS which results in an adaptive dilation and shrinkage of the windows depending on the temporal extent of the underlying PQSS.

In [2, 3], Svendsen et. al. proposed a ML segmentation algorithm for speech signals. Their algorithm uses a single £xed window size for speech analysis and then clusters the frames which are spectrally similar for sub-word unit design. We emphasize that this is different from our technique where we use variable sized windows to achieve the objective of piecewise quasi-stationary spectral analysis. Recently, Achan et. al.[5] have proposed a segmental HMM for speech waveforms which identi£es waveform samples at the boundaries between glottal pulse periods with applications in pitch estimation and time-scale modi£cations.

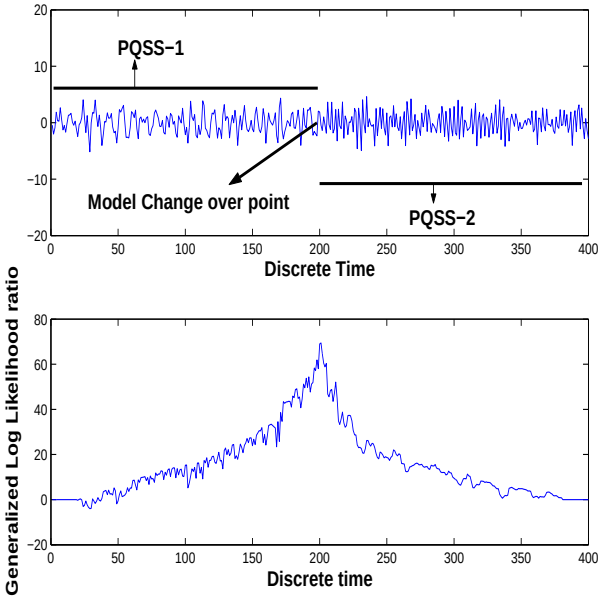

Figure 1: *Exact detection of the model change over point for a 6th order AR Gaussian process using Generalized log likelihood ratio test (GLRT).*

This paper is divided into £ve sections. In Section 2, we formulate the ML detection problem for identifying the transition from one PQSS to another. In Section 3, we apply the proposed algorithm to the real speech signals. 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, $x[n], n \in [1, N]$ whose generative LP £lter parameters change from $\mathbf{A}^1 = [a^1(1), a^1(2)....a^1(p)]$ to $\mathbf{A}^2 = [a^2(1), a^2(2)....a^2(p)]$ at time n_0 where $n_0 \in [1, N]$. The excitation signal power can also change from σ_1 to σ_2 The general form of the PSD of this signal is well known to be,

$$P_{xx}(f) = \frac{\sigma_u^2}{|1 - \sum_{p=1}^{P} a(p) \exp(-j2\pi pf)|^2} \tag{1}$$

The hypothesis test consists of:

- $\mathbf{H}_0$: No change in the PSD of the signal $x(n)$ during $n \in [1, N]$, LP £lter parameters are $\mathbf{A}^0$=$\mathbf{A}^1$ and the excitation(residual) signal power is $\sigma_0 = \sigma_1$.
- $\mathbf{H}_1$: Change in the PSD of the signal $x(n)$ at n_0, where $n_0 \in [1, N]$, LP £lter 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 £lter parameters and $\hat{\sigma}_0$ denote the MLE of the residual signal power under the hypothesis $\mathbf{H}_0$. The MLE estimate of the £lter parameters is equal to their MMSE estimate due to the Gaussian distribution and hence can be computed by the Levinson Durbin algorithm [1]. Let $\mathbf{x}_1$ denote $[x(1), x(2), ...x(n_0)]$ and $\mathbf{x}_2$ denote $[x(n_0 + 1), ...x(N)]$. Under hypothesis $\mathbf{H}_1$, ($\hat{\mathbf{A}}^1$, $\hat{\sigma}_1$) are the MLE of ($\mathbf{A}^1$, σ_1) based on $\mathbf{x}_1$, and ($\hat{\mathbf{A}}^2$, $\hat{\sigma}_2$) are the MLE of ($\mathbf{A}^2$, σ_2) based on $\mathbf{x}_2$. A generalized likelihood ratio test (GLRT)[6] would decide $\mathbf{H}_1$ if,

$$L(\mathbf{x}) = \frac{p(\mathbf{x}_1|\hat{\mathbf{A}}^1, \hat{\sigma}_1)p(\mathbf{x}_2|\hat{\mathbf{A}}^2, \hat{\sigma}_2)}{p(\mathbf{x}|\hat{\mathbf{A}}^0, \hat{\sigma}_0)} > \gamma \tag{2}$$

We note that the total number of samples in $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same as in $\mathbf{x}_0$. Therefore, their likelihoods can be compared directly in (2). Usually, the threshold γ is experimentally tuned. 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 [6]. Under $\mathbf{H}_1$, we assume that there are two distinct PQSS, 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 PQSS. 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 £lter parameters $\hat{\mathbf{A}}^0$. Therefore,

$$p(\mathbf{x}|\hat{\mathbf{A}}^0, \hat{\sigma}_0) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{N/2}} exp(\frac{-1}{2\hat{\sigma}_0^2}\sum_{n=1}^{N}(e_0^2(n)))$$
$$\text{where } e_0(n) \text{ is the residual error and,} \tag{3}$$
$$e_0(n) = x(n) - \sum_{i=1}^{p} a^0(i)x(n-i), \;\; n \in [1, N]$$
$$\text{and } \hat{\sigma}_0^2 = \frac{1}{N}\sum_{n=1}^{N} e_0^2(n))$$

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 probabilities of corresponding residual signals whose functional forms are similar to the ones in (3). Substituting these expressions in (2), it can be simpli£ed to,

$$L(\mathbf{x}) = \frac{\hat{\sigma}_0^{N/2}}{\hat{\sigma}_1^{n_0/2}\hat{\sigma}_2^{(N-n_0)/2}} \tag{4}$$

In the present form, the GLRT L(**x**) has a natural interpretation which is as follows: If there is indeed a change point in the segment **x** then it has $2P$ degrees of freedom. Under $\mathbf{H}_0$, we encode **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 $\mathbf{H}_1$, we use $2P$ degrees of freedom to encode **x**. Therefore, the coding (residual) errors $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be minimized to the minimum possible.[3] This will result in $L(\mathbf{x}) > 1$. On the other hand, if there is no change point in the segment **x** then it can be shown that for large n_0 and N, the coding errors are all equal ($\hat{\sigma}_0^2 = \hat{\sigma}_1^2 = \hat{\sigma}_2^2$). This will result in $L(\mathbf{x}) \simeq 1$.

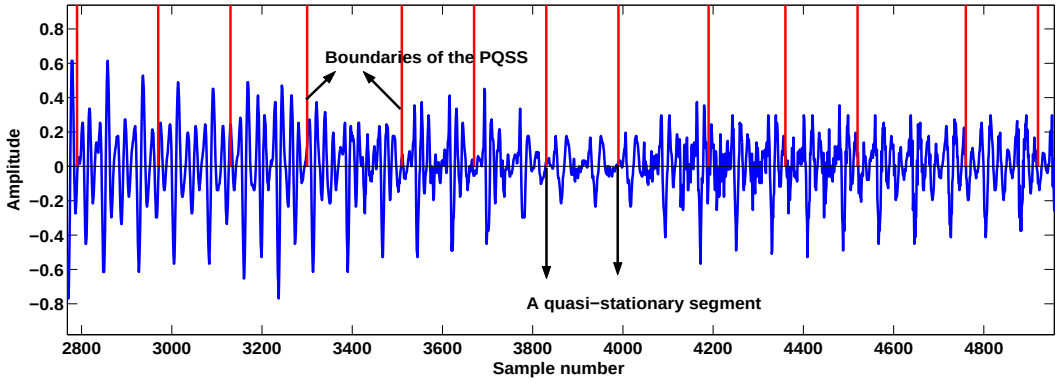

Figure 2: Piece-wise quasi stationary segments (PQSS) of a speech signal as detected by the algorithm with $\gamma = 3$ and LP order $p = 14$.

An example is illustrated in £gure 1. The top pane shows an instance of a 6^{th} order AR Gaussian process which has a model change point at $n_0 = 200$. In the bottom pane, we plot the GLRT as the function of the hypothesized change over

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

point n. The GLRT achieves the maximum at $n_0 = 200$ which is indeed the PSD change over point.

3 Multi-scale Piecewise stationary analysis of speech signals

We have used the GLRT $L(\mathbf{x})$ (4) to facilitate piecewise quasi-stationary analysis of speech signals. The actual algorithm used is outlined below,

Given signal [x(0),x(1),....x(N)], consider
two segments, LeftSegment=[x(L_s)...x(L_e)]
and RightSegment=[x(R_s),...x(R_e)].
Detect if model changes at $R_s = L_e + 1$.
1. INITIALIZATION:
$L_s = 0$, $L_e = L_s$ + LEFTMIN,
$R_s = L_e + 1$, $R_e = R_s$ + MINRIGHT.

2. Evaluate GLRT with the current
boundaries of the two segments.

3. If GLRT < γ, no model change point at L_e. Set L_e=L_e+INCR,
$R_s = L_e + 1$, $R_e = R_s$ + MINRIGHT.

4. If GLRT >= γ, model change point at L_e.
Set L_s=L_e, L_e=L_s + LEFTMIN, $R_s = L_e + 1$, $R_e = R_s$ + MINRIGHT.

5. If R_e < N , go to (2).

We constrain the minimum duration of the two analysis windows to be MINLEFT=10ms and MINRIGHT=5ms. This is done in order to avoid estimating LP parameters from a very small number of speech samples. If there is no model change point detected at the current boundary, then the duration of the left segment is incremented by INCR=1.25ms. When a model change point is detected, the left segment is considered a PQSS. All the speech samples in this PQSS are windowed together to obtain the corresponding Mel-frequency cepstral coef£cients (MFCC)[10]. We emphasize that the resulting PQSSs are not constrained to be of equal size which explains the use of the term "multi scale" in our algorithm. In fact, the use of the threshold γ can be avoided by searching for a local maxima of the GLRT and assigning it as a model change point. This is evident in the £gure 1 where the GLRT achieves a distinct maximum at the model change point. Ajmera et. al.[4] have successfully used this technique for speaker change detection. However, in our algorithm, we have used an explicit threshold $\gamma = 3$. This threshold was obtained by a visual inspection of the quasi-stationarity of the segmented speech signal as returned by the algorithm. In £gure 3, we show the boundaries of the PQSS as detected by the algorithm with $\gamma = 3$. As, we found this segmentation to be reasonably quasi-stationary, we adopted the threshold value $\gamma = 3$ for all the experiments reported in this paper. In the future work, we will incorporate the

maxima detection as a model change point avoiding the use of the threshold. However, it is worth noting that the use of a threshold is quite usual in the detection theory[6]

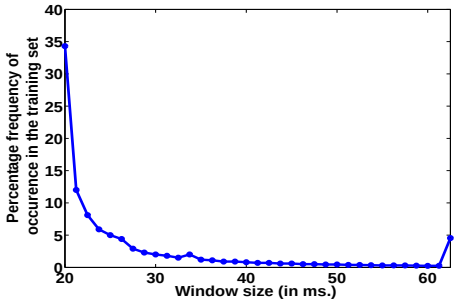

Figure 3: *Distribution of the PQSS window sizes detected and then used in the training set*

4 Experiments and Results

In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the OGI Numbers corpus [9]. It contains spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words. We used the algorithm in the section (3) to detect PQSS. The computation of a MFCC feature vector from a very small segment (such as 10ms) is inherently very noisy.[4] Therefore, the duration of a PQSS as detected by the algorithm was constrained to be in the interval $[20ms, 62.5ms]$. A £xed LP order $p = 14$ and the threshold $\gamma = 3$ in (2) was used in the algorithm. The value of the threshold $\gamma = 3$ resulted in a reasonable segmentation of the speech signal in terms of the PQSS. Figure 3 shows a speech waveform with the overlaid boundaries of the detected PQSS using $\gamma = 3$. The distribution of the duration of the PQSSs is shown in £gure 3. Nearly 35% 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 $30ms - 62.5ms$. Throughout the experiments, Mel-frequency cepstral coef£cients (MFCC) [10] and their temporal derivatives have been used as speech features. Four feature sets were generated:

1. [MFCC+Deltas:] 39 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coef£cient) with cepstral mean subtraction and their standard delta and acceleration features. Spectrum computation over a window of length 20ms and a shift of 12.5ms.

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

2. [MFCC+Deltas:] Same as above except that the spectrum is computed over a window of length 50ms.

3. [Concatenated MFCC+Deltas:] 78 element feature vector which is a concatenation of the above two feature vectors.

4. [Multi-scale PQSS MFCC+Deltas:] 39 element feature vector consisting of 13 MFCCs and deltas. For a given frame, the window size is dynamically chosen using the proposed algorithm ensuring that the windowed segment is quasi-stationary. A constant shift size of 12.5ms same as in baseline.

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK [8] on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. Although the multi-scale PQSS features were analyzed using variable sized windows in the range $[20ms, 62.5ms]$, a constant shift size of $12.5ms$ was used as in the usual baseline features. This was done for the sake of simplicity in training a £xed HMM topology of 3 states per triphone as in the baseline system. We believe that this may be a limitation. In future experiments we will perform a cross-validation of the optimal number of states per triphone depending on the distribution of its PQSS. The speech recognition results in clean conditions for the two £xed scale baseline, concatenated multi-scale and proposed multi-scale systems are given in table 1. The proposed multi-scale system has a 5.1% WER. This corresponds to a relative improvement of 10% WER over the two baselines with WERs 5.9% and 5.8%. The concatenation of MFCC feature vectors derived from 20ms and 50ms long windows has a 5.7% WER. The slight improvement in this case may be due to the multiple scale information present in this feature. It is worth noting that the concatenated MFCC vector based ASR system has twice the number of HMM-GMM parameters as compared to the proposed system. The proposed multi-scale PQSS based MFCC system has a 8% relative improvement over the multi-scale concatenated system.

Table 1: *Word error rate in clean conditions*

MFCC 20ms	5.8
MFCC 50ms	5.9
Concat. MFCC (20ms, 50ms)	5.7
Proposed Multi-scale PQSS MFCC	5.1

5 Conclusion

We have proposed a novel criterion for multi-scale piecewise quasi-stationary analysis of speech signal. This technique overcomes the limitations of a single

£xed scale spectral analysis techniques which blindly assumes each analysis window to be stationary. The proposed piecewise quasi-stationary analysis technique yields 10% to 8% relative improvement over the single £xed scale and concatenated multiple but £xed scale analysis techniques. In future work, we will explore the possibility of detecting the PQSS using the local maxima of the generalized likelihood ratio and the use of a non-uniform HMM-topology.

6 Acknowledgments

This work was supported by European Commission 6th Framework Program project DIVINES under the contract number FP6-002034. The £rst author would like to thank Dr. Iain Mccowan, Prof. Raymond Knopp, Dr. Jonas Samuelsson and Prof. Bastiaan Kleijn for valuable discussions with them which helped this work.

References

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

[2] T. Svendsen, " On the Automatic Segmentation of speech signals, " Proc. of IEEE ICASSP, 1987.

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

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

[5] 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.

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

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

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

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

[10] 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.