IDIAP RESEARCH REPORT

On Multi-scale Fourier Transform Analysis of Speech Signals

Vivek Tyagi [a] Hervé Bourlard [a,b]

IDIAP–RR 03-32

June 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, Martigny, Switzerland
[b] EPFL, Lausanne, Switzerland

IDIAP Research Report 03-32

On Multi-scale Fourier Transform Analysis of Speech Signals

Vivek Tyagi Hervé Bourlard

June 2003

Abstract. In this paper, we introduce a novel algorithm to perform multi-scale Fourier transform analysis of piecewise stationary signals with application to automatic speech recognition. Such signals are composed of quasi-stationary segments of variable lengths. Therefore, in the proposed algorithm, signals are analyzed with multiple-sized windows. Resulting power spectra are then normalized such that they all have unit energy, followed by entropy computation of each power spectrum. These entropies are further normalized because they are computed over different number of sample points. Amongst these power spectra, the one with the minimum normalized entropy is retained as optimal power spectrum estimate. In experiments with speech signals, it is shown that the proposed multi-scale Fourier transform based features yield an increase in speech recognition performance in various non-stationary noise conditions when compared directly to single fixed scale Fourier transform based features.

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 scale. There is another subtle point which arises in spectrum estimation of a signal using fixed scale discrete Fourier transform (DFT) analysis. The DFT computation of a signal implicitly assumes that the signal is periodic with the base period of the analysis window. If the original period of the signal is much smaller than the length of the analysis window, the resulting power spectrum estimate is reasonably accurate. However, if the length of the analysis window is comparable to the period of the signal, the short time power spectrum obtained is a poor estimate of the signal spectrum [1].

In this work, we use multiple window sizes and compute the corresponding power spectral density (PSD) for each window size. These PSDs are then normalized such that they all have equal energies. This can also be interpreted as transforming PSDs to probability mass functions (PMF). Amongst these normalized PSDs, the one with the minimum spread of the energy across the frequency axis is used as a spectral estimate of the quasi-stationary segment under analysis. For a particular window size, the spread of the energy across the frequency axis is measured in terms of the entropy of the corresponding PMF.

This paper is divided into five sections. In Section 2, we discuss the limitations of single scale Fourier transform analysis of signals. In Section 3, we describe the proposed algorithm for the multi-scale Fourier transform analysis. The experimental setup and the results are described in Section 4.

2 Limitations of Single-scale Spectrum Analysis

Speech signals can be assumed to be composed of piecewise stationary segments (PSSs) of variable lengths.

However, most speech signal feature extraction modules analyze the incoming signal into frames of equal lengths, followed by discrete Fourier transform (DFT) operation on each frame. An inherent problem with this analysis is that a single frame size is not optimal for variable sized PSSs which form the signal. Optimal window size for a particular PSS is the biggest size without leakage from neighbouring PSS. As an example, let us assume $s[n]$ is a speech signal composed of PSSs $x_1(n)$, $x_2(n)$,..... $x_N(n)$

$$\begin{aligned} s[n] &= x_1(n) \;\; n \in [0,\, 2L-1] \\ &= x_2(n-2L) \;\; n \in [2L,\, 3L-1] \\ &= .... \end{aligned} \tag{1}$$

Consider three rectangular windows w_1, w_2, w_3 of sizes L, 2L, 3L respectively. Let F_i be the magnitude of the discrete time Fourier transform (DTFT) of signal $s[n]$ windowed by w_i and W_i be the DTFT of w_i. X_i is the DTFT of x_i. Then it follows:

$$F_i(e^{j\omega}) = \mid X_1(e^{j\omega}) * W_i(e^{j\omega}) \mid^2, \quad i \in [1,2] \tag{2}$$

$$F_3(e^{j\omega}) = \mid (X_1(e^{j\omega}) + X_2(e^{j\omega})e^{-j\omega 2L}) * W_3(e^{j\omega}) \mid^2 \tag{3}$$

W_i has a spectrum similar to that of a narrow band-pass filter. The Bandwidth of the filter decreases with increasing window length. Therefore, F_2 is a better spectrum estimate than F_1 because it has finer frequency resolution. Although W_3 has the finest frequency resolution, from a pattern recognition point of view, F_3 is a poor spectrum estimate because it is a mixture of two PSS spectra.

Instances of feature vectors like F_3 decrease the discrimination between various PSS. Let $P_i(e^{j\omega})$ be normalized $F_i(e^{j\omega})$ such that,

$$P_i(e^{j\omega}) = F_i(e^{j\omega}) / \int_{-\pi}^{\pi} F_i(e^{j\omega}) d\omega \tag{4}$$

We note that $P_1(e^{j\omega})$ and $P_2(e^{j\omega})$ have the same functional form but different parameters. Due to the fact that W_2 has a narrower pass-band than W_1, the convolution operation in (2) results in $P_2(e^{j\omega})$ having less variance than $P_1(e^{j\omega})$. Therefore, it can be shown that the entropy of $P_2(e^{j\omega})$ is less than that of $P_1(e^{j\omega})$. Moreover, if the distribution of the spectral power in $X_1(e^{j\omega})$ is significantly different than that in $X_2(e^{j\omega})$, it would imply that the spread of the spectral power in $F_3(e^{j\omega})$ is more than that in $F_2(e^{j\omega})$. Therefore, it can be safely assumed that the entropy of $P_2(e^{j\omega})$ will be less than entropy of $P_3(e^{j\omega})$.

The above example shows that under suitable conditions, the spectrum estimate with minimum spectral power spread is the optimal estimate amongst various spectra, computed over multiple window sizes. In this work, spectral power spread is measured in terms of the entropy of the normalized spectrum. The required suitable conditions are that:

1. The signal can be assumed to be composed of PSSs.
2. Wide range of window sizes are required, to be able to segment the incoming signals into valid PSSs.

It is well known that the power spectral density of a windowed signal is the DTFT of the temporal autocorrelation of the windowed signal. For a particular PSS, a window size determines the number of samples over which the temporal autocorrelation is performed. The larger the number of terms, the better the autocorrelation estimate and therefore the better the spectral estimate. However, if the window length is larger than the length of the PSS, terms from adjoining PSS start to enter the autocorrelation estimate. Consequently, the spectrum has contributions from more than one PSS leading to poor discrimination ability in pattern recognition problems. Multi-scale analysis alleviates this problem by choosing the optimal scale for the Fourier transform analysis of piecewise stationary signals.

3 Multi Scale Spectral Analysis

Multi scale signal processing involves performing DFT analysis of a non-stationary signal (which can be assumed to consist of PSSs) with several variable sized windows. Let $w_{1\ldots M}$ be the windows and $L_{1\ldots M}$ be the corresponding window size such that $L_1 < L_2 < \ldots\ldots < L_M$. Let X_i be the DFT of $x(n)$ multiplied with window w_i .

$$\begin{array}{c} X_i(k) = \sum_{n=0}^{L_i-1} x(n) w_i(n) \exp(\frac{-j2\pi nk}{L_i}) \ , \\ \forall\, k \ \in [0,\ L_i - 1] \ and \,\forall\, i \ \in [1,\ M] \end{array} \tag{5}$$

For a given window size, if the window has at least one base period of the quasi-stationary segment in it, and it does not capture any other adjoining quasi-stationary segments, then the power spectrum obtained will be a good estimate of the actual spectrum. Otherwise, the resulting power spectrum will have poor time and frequency resolution. We use this hypothesis to automatically choose an optimal window size for a given quasi-stationary segment under analysis. We take windows of different sizes and compute their corresponding power spectrum. The resulting power spectrum is normalized so that it takes the form of a PMF. Let P_i be the PMF obtained by normalizing $\mid X_i \mid^2$.

$$\begin{array}{c} P_i(k) = \mid X_i(k) \mid^2 / \sum_{j=0}^{L_i-1} \mid X_i(j) \mid^2 \ , \\ \\ \forall\, k \ \in [0,\ L_i - 1] \ and \,\forall\, i \ \in [1,\ M] \end{array} \tag{6}$$

Next, we compute the entropy of each PMF resulting from several analysis windows and again normalize them to the scale (0, 1). Let H_i be the entropy of PMF P_i and H_i^{norm} be the normalized entropy.

$$H_i = -\sum_{k=0}^{L_i-1} P_i(k)\ln(P_i(k))\ \forall\, i\ \in [1,\ M] \tag{7}$$

The maximum entropy of a PMF defined over L points is $\ln(L)$. Therefore normalizing H_i by $\ln(L_i)$ ensures that all the entropies lie in the interval (0,1).

$$H_i^{norm} = H_i/\ln(L_i) \tag{8}$$

The window size which gives the minimum entropy is chosen to be the optimal window size for the quasi-stationary segment under analysis.

$$H_m^{norm} = \operatorname*{argmin}_i\ H_i^{norm}\ \forall\, i \in [1,\ M] \tag{9}$$

and $\mid X_m(k) \mid^2$ is retained as the power spectral density estimate of this segment. The signal is shifted by the last optimal window length and the algorithm is repeated for the new quasi-stationary segment.

Given a segment which is quasi-stationary for a duration greater than the largest window size L_M, the above mentioned algorithm will select X_M as the spectral estimate as it will have the minimum entropy. With the increasing window size, the main-lobe of the window spectrum will decrease. This will decrease the smearing effect due to convolution, giving sharper frequency resolution while monotonically decreasing the entropy in (8). On the other hand, if a segment is quasi-stationary for a duration L such that $L_i < L < L_{i+1}$ then the algorithm will select X_i as the spectral estimate. The windows W_{i+1}, W_{i+2}, ..., W_M will capture more than one quasi-stationary segment in the analysis, leading to blurring across time and frequency. Therefore, the corresponding entropy in (8) will be high. The windows W_1, W_2, ..., W_i will capture only one quasi-stationary segment in the analysis with increasing frequency resolution and monotonically decreasing the entropy. Therefore, the proposed algorithm will select the largest possible window size amongst the given sizes, such that the signal remains quasi-stationary within the window duration.

4 Experiments and Results

In order to assess the effectiveness of the proposed multi-scale signal processing, speech recognition experiments were conducted on the Numbers corpus [4]. We have used a much simpler version of the proposed algorithm, where incoming speech signal is analyzed with only two window sizes (12.5ms, 37.5ms). For reasons of simplicity in statistical modeling using Hidden Markov Models and Gaussian Mixture Models (HMM-GMM), the frame shift is kept constant at 12.5ms rather than using the variable shift rule proposed in the algorithm. Throughout the experiments, Mel-frequency cepstral coefficients (MFCC) [2] and their temporal derivatives have been used as speech features. Three feature sets were generated:

1. [MFCC+Deltas:] 39 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) with cepstral mean subtraction and their standard delta and acceleration features. Spectrum computation over a Hamming window of length 32ms.

2. [Concatenated MFCC+Deltas:] 78 element feature vector which is concatenation of two 39 element feature vectors. These two vectors are derived using windows of lengths 12.5ms and 37.5ms respectively. Individually, they consist of 13 MFCCs (including 0^{th} cepstral coefficient) with cepstral mean subtraction and their standard delta and acceleration features.

3. [Multi-scale MFCC+Deltas:] 39 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) with cepstral mean subtraction and their standard delta and acceleration features. Two Hamming window sizes of lengths 12.5ms and 37.5ms were used. For a given frame, the size was chosen dynamically using (9).

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK [3] 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.

To verify the robustness of the features to noise, the clean test utterances were corrupted using Factory and Car noises from the Noisex92 [5] database. The speech recognition results for the fixed scale baseline, concatenated multi-scale and proposed multi-scale systems, in various levels of noise are given in Tables 1 and 2. The proposed multi-scale system performs better than the concatenated multi-scale system in the noisy conditions and does so with half the number of parameters of the concatenated system.

In these experiments, there are two opposing factors at work. The assumption of speech signal being composed of PSSs is not always correct. Therefore, the optimality criterion in (9) does not hold true for all the cases. As a result, instances of a particular quasi-PSS might get analyzed by two or more different window sizes. This leads to a mismatch between the training and the testing conditions. On the contrary, this problem is inherently absent in the other two systems due to the use of the fixed scale analysis. This explains the slight degradation of the proposed multi-scale system in the clean conditions. However, in the case of segments where the optimality criterion in (9) holds true, we get the advantage of better time and frequency resolution. This advantage is more evident in noisy environments, where the enhanced time and frequency resolution more than compensates for the mismatch problem discussed above.

Table 1: *Word error rate results for factory noise*

SNR	MFCC	Concatenated MFCC	Multi-scale MFCC
Clean	6.6	6.4	7.3
12 dB	23.1	22.4	19.5
6 dB	48.8	50.0	40.7

Table 2: *Word error rate results for car noise*

SNR	MFCC	Concatenated MFCC	Multi-scale MFCC
Clean	6.6	6.4	7.3
12 dB	18.2	16.4	15.7
6 dB	38.1	35.1	32.4

5 Conclusion

We have shown that the multi-scale Fourier transform analysis can yield a better estimate of the speech signal spectrum. The presented algorithm, automatically finds the optimal size of the window for the quasi-stationary segment under analysis, achieving optimal time and frequency resolution for such a segment. To have a rigorous comparison of the proposed multi-scale processing with other such techniques, a concatenated multiple scale feature vector based system was trained. The concatenated feature vectors were twice the size of the proposed processing based feature vectors, with twice the number of parameters in HMM-GMM system as in the proposed system. In the clean case, the proposed system performed poorer, which can be attributed to the mismatch phenomenon explained in the previous section. In the noisy conditions, the proposed feature extraction approach performed better than the other two systems.

References

[1] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, N.J., USA, 1989.

[2] 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.

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

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

[5] A. Varga, H. Steeneken, M. Tomlinson and D. Jones, " The NOISEX-92 study on the effect of additive noise on automatic speech recognition, " Technical report, DRA Speech Research Unit, Malvern, England, 1992.