ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

EARLYZER : PERCEPTUALY MOTIVATED ROBUST TFR OF SPEECH *

Avadhanulu, J.V[1]. Mathew, M. and Sreenivas, T.V.
Department of Electrical Communication Engineering
Indian Institute of Science
Bangalore 560 012, India. (Email: tvsree@ece.iisc.ernet.in)

Abstract: *Development of robust and efficient front-end is crucial for robust ASR. Proper time and frequency resolution of the TFR of speech, motivated by the auditory models is considered an important factor for robustness. An efficient method of realizing a variable resolution TFR using DTFT/Goertzel algorithm is proposed instead of the standard FFT based approach. It is shown that the new representation, called EarLyzer, is more robust than the FFT based Mel frequency cepstral coefficient representation for an automobile noisy speech recognition task.*

1. INTRODUCTION

Time Frequency Representation (TFR) of speech holds the key to good performance in Automatic Speech Recognition (ASR): information necessary for speech recognition is to be *extracted* while *discarding* components due to noise and speaker/speaking variabilities. Optimum resolution in TFR of speech is an open problem that has attracted much attention [1]. Optimum resolution would also determine the robustness of ASR performance. The commonly used TFR is to simulate Mel/Bark scale filter-bank using a DFT representation. Another approach is to use detailed simulation of auditory filter properties using explicit FIR/IIR functions, such as Gamma-tone filter-bank. For ASR applications, auditory filter-banks should also be efficiently realizable, which is not true with the latter approach. While the DFT based Mel filter-bank is efficiently implemented, it lacks certain auditory properties, such as variable temporal resolution. EarLyzer is an approach to obtain variable time and frequency resolutions, while retaining the computational efficiency.

EarLyzer is a Mel-spaced overlapping filter bank, which uses DTFT to compute at the exact frequencies on the Mel-scale. The desired frequency response has the conflicting requirements of overlap between adjacent channels and high attenuation in farther channels. This is achieved by the design of analysis windows for each of the channels in conjunction with multi-resolution analysis followed by 2-D smoothing and decimation. EarLyzer is realized as a bank of filters with P=(2Q+1) channels to produce a Q channel output after 2-D filtering and decimation by a factor of 2 in both the time and frequency domains. Thus, the basic analysis filter bank has twice the number of channels, each of which operates at twice the feature rate finally required for the application. The EarLyzer spectrum is further processed by a noise compensation block wherein the reference pattern is compensated for the noise present in the test pattern, to provide robust pattern matching. This is unlike the usual spectral subtraction which leads to inconsistency of the estimated spectral density.

2. FILTERBANK THROUGH DTFT

The overall signal processing chain is shown schematically in Fig. 1. Let Q be the number of channels required at the output of EarLyzer. We select P =2Q+1 channels in the DTFT based filter bank. The output of i^{th} channel evaluated through DTFT is given by:

$$X(\omega_i, n) = X_{STFT}(e^{j\omega}, n)\Big|_{\omega=\omega_i} \quad 1 \le i \le P, \forall n,$$
$$= \sum_{m=0}^{L_i-1} x(n-m)h_i(m)e^{-(j2\pi f_i / f_s)m} \qquad (1)$$

where $h_i(m)$ is the window sequence for the i^{th} channel, f_i is the Mel scale center frequency of the i^{th} channel and f_s is the sampling frequency. In (1), a finite window of length L_i is used and the X_{STFT} is evaluated at P Mel-spaced frequencies using the Goertzel's algorithm, which is computationally efficient. The Mel scale frequencies are given by [2]

$$f_{mel} = 2590 \log_{10}(1 + f/700). \qquad (2)$$

The Mel-scale spacing is approximately linear in the frequency range up to 1 kHz and logarithmic at higher frequencies. The mapping relates acoustic frequency to perceptual frequency and the human auditory system integrates the frequency regions in bands called critical bands. A critical band filter can be considered as a band pass filter whose frequency response corresponds to the frequency selectivity of basilar membrane and the auditory neurons. For speech, the frequency range of interest is approximately 100 Hz to 3kHz, which is covered by 19 critical bands, giving rise to P=39 channels in the DTFT based filter bank. The center frequencies f_i of the 39 channels are shown in Fig. 2(a).

We define the bandwidth of the i^{th} filter as

$$b_i^{rect} = f_{i+1} - f_{i-1} \qquad (4)$$

This bandwidth for the 39 channels is shown in Fig. 2(b). It is evident from (1) that if we choose rectangular windows with $L_i=f_s/b_i$, the main lobe nulls in the frequency response of i^{th} channel will be located at f_{i+1} and f_{i-1}. Instead, we choose Hamming window function

* The work is supported by a research project sponsored by M/s Ericsson Inc., USA.
[1] Currently with M/s Sigmatech, Bangalore.

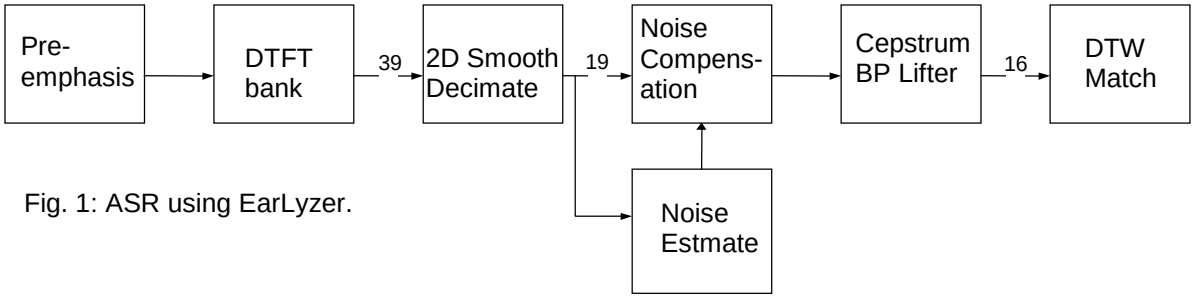


Fig. 1: ASR using EarLyzer.

which has better reduction of sidelobe levels and attenuation in farther channels. However, this increases the bandwidth to $b_i^{hamm} = 2(f_{i+1} - f_{i-1})$ and hence we choose,

$$L_i^{hamm} = Int(2 f_s / b_i^{hamm}) \quad (5)$$

where Int(.) denotes the nearest integer. The window functions are shown in Fig. 2(c). In the DFT based Mel filter-bank, the window width is deliberately chosen higher to enable finer frequency resolution; the successive frame overlap is fixed as 50% or 75%. Such an analysis results in a uniformly poor temporal resolution for all channels and a coarse sampling of wide-band channels. Instead, in the EarLyzer, the frame shift is equal to min{L_i}, which is independent of the resolution of the narrow-band filters. This assures good temporal resolution of the high frequency bands, independent of the low frequency bands. However, to reduce the overall TFR size (hence, pattern matching complexity) the P channel output is convolved with a 3x3 rectangular smoothing window and then decimated by a factor of 2 in both time and frequency dimensions. For f_s=8kHz, it is found that 39 channel output results in 166 frames/sec which yields a 19 channel output at a rate of 83 vectors/sec after smoothing and decimation. It may be noted that in this approach, the overlap between successive frames increases for lower frequency channels and is zero for the 39th channel (see Fig. 2(d)). The frequency response of the DTFT analysis is shown Fig. 3 along with that after smoothing and decimation, for the first ten channels; the flat frequency response in this is noteworthy. The maximum ripple of 0.65 dB (which occurs at the crossover point) may be compared with 3 dB in Mel filter banks. The significance of this can be seen as follows. Consider a 'glide' that moves across the filter channels with equal loudness. For this the Mel filter bank introduces an amplitude modulation of 3 dB where as EarLyzer introduces an amplitude modulation of only 0.65 dB. This reduction of cross coupling between amplitude modulation and frequency modulation would cause reduction of artifacts in the speech TFR which in turn would improve the speech recognition performance.

3. NOISE COMPENSATION

Human auditory perception is known to be adaptive to ambient noise. We incorporate a noise compensation block to further improve the robustness of the EarLyzer TFR. It is assumed that noise is slowly time-varying and a separate voice-activity-dector (VAD) is available. filter Based on this an adaptive estimate of the noise psd is obtained upto the start of the speech segment. Thus,

$$\bar{N}(\omega_i,n) = \alpha_i \bar{N}(\omega_i,n-1) + (1-\alpha_i)|X(\omega_i,n-1)|^2, \; n < n_{VAD} \quad (6)$$

where $\alpha_i = e^{-2\pi CB_i}$ (7)

and B_i is the normalized bandwidth given by $b_i/\max(b_i)$ and C is a suitable constant. The averaging time constant of each of the filters is 'matched' to the bandwidth of the filter, as may be inferred from the knowledge of the human auditory models.

For noise compensation, the usual algorithm is to use the time averaged estimate of the noise spectrum in a non-linear spectral subtraction (NSS) [3] given by

$$S(\omega_i, n) = \Gamma\{X(\omega_i, n) - \bar{N}(\omega_i, n)\}, \; n > n_{VAD} \quad (8)$$

where Γ is a non-linear function that prevents the psd from taking negative values. Clearly NSS gives rise to errors in the estimated spectrum whenever the noise spectrum exceeds the signal spectrum. This leads to further difficulties when we choose logarithmic compression in downstream processing. These discontinuities in the noise subtracted spectrum can be avoided by compensating the clean training data instead of the noisy test data. i.e.

$$\tilde{R}(\omega_i,n) = R(\omega_i,n) + \beta \bar{N}(\omega_i, n_{VAD}) \; . \quad (9)$$

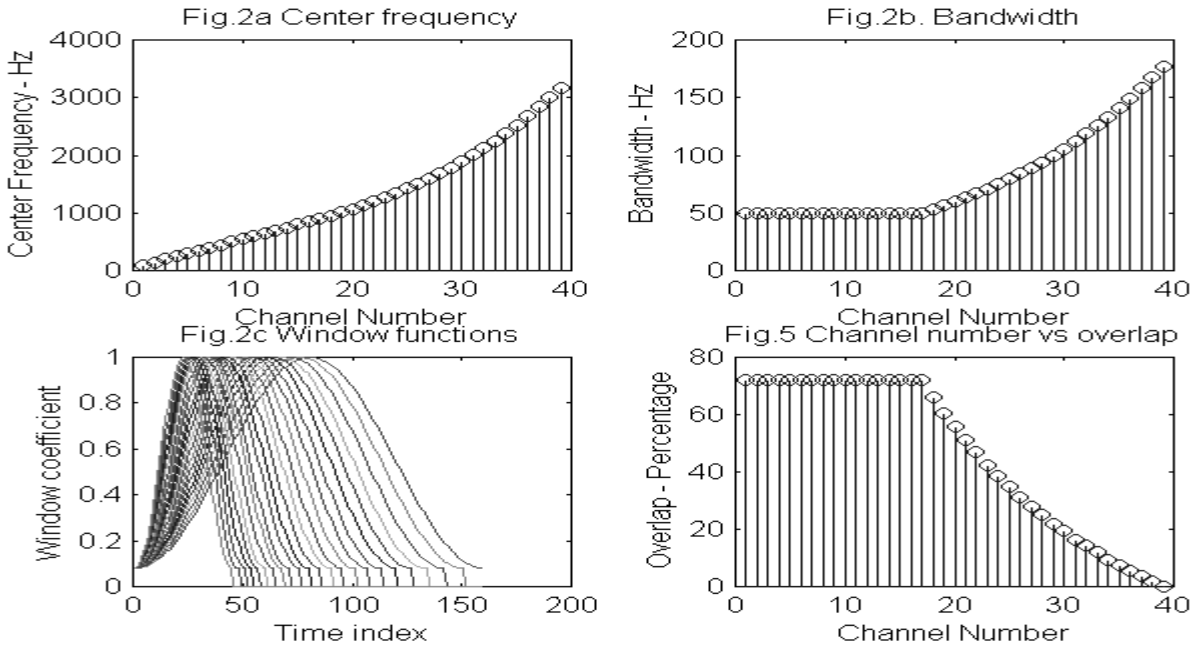
Here R and T denote the reference and test TFRs, respectively, and β is a scale factor chosen such that

$$Avg\text{–}seg\text{–}SNR(R) = Avg\text{–}seg\text{–}SNR(T) \quad (10)$$

Fig. 4(a) shows the waveform of the utterance "John Smith" spoken by an adult male, sampled at 8kHz spoken in a quiet environment. Fig. 4(b) shows the same word spoken by the same speaker in a moving automobile. Fig. 4(c) shows the contour plot of the EarLyzer TFR of the noisy speech and Fig. 4(d) shows the contour plot of the EarLyzer TFR of the noise compensated clean speech. The similarity between Fig4(c) and Fig 4(d) is evident even without any dynamic range compression.

4. SPEECH RECOGNITION EXPERIMENT

The EarLyzer is evaluated on a speaker dependent, isolated word recognition task using DTW pattern matching. From the EarLyzer output of 19 channels, 16 band pass liftered cepstral coefficients are derived at the rate of 83 vectors/sec. Euclidean distance measure is used for pattern matching. Reference patterns are



generated from the clean speech recorded in a parked car. Test patterns are generated from the noisy speech in a moving car, at an Avg-seg-SNR of ~0dB. A database consisting of 10 speakers (5 male and 5 female) each with a vocabulary of 30 proper nouns, is used in the experimental evaluation. The database was manually end pointed for VAD.

The EarLyzer is evaluated without the noise compensation block and the results are compared with those obtained from FFT based Mel Filter-bank; these are shown in Table-1. It can be seen that the EarLyzer out performs the FFT based Mel filter-bank. Next, the EarLyzer is evaluated with the same vocabulary and same speakers but under 5 different kinds of automobile noise conditions, to establish the performance obtainable in real world conditions; these results are presented in Table-2. It can be seen that under all noise conditions, the EarLyzer out performs the MFCC even at 100 vec/sec representation. (In these tables, E denotes the EarLyzer algorithm and M denotes the FFT based Mel Filter bank algorithm. The subscript denotes the feature vector rate in vectors/second and the letters in the brackets indicate other signal processing incorporated. viz., NC indicates the reference vectors have been noise compensated, and D indicates that dynamic features have been used in pattern matching.) Table-3 shows the correct class recognition of the 300 patterns (30x10 spkrs) along with recognition as 2nd best, 3rd best, etc. These results are compared between three different algorithms for one noise condition, i.e., automobile-2. It can be seen that even without noise compensation, EarLyzer performs better than the MFCC.

Table-1: EarLyzer Vs MFCC @ 83 vec/sec

Algorithm	*Correct word recognition*
M_{83}	88.33% (265/300)
E_{83}	90.66% (272/300)

Table-2: Recognition score in different noise conditions.

Automobile type	$E_{83}(NC)$	M_{100}	$M_{100}(D)$
1	93.33%	91.33%	97%
2	98.00%	90%	97%
3	82.66%	71.66%	81%
4	95.66%	95.66%	96.33
5	80.5%	64.66%	71.66%

Table-3: Correct word as ‘k’ nearest neighbor

Rank-k	M_{83}	E_{83}	$E_{83}(NC)$
1	265	272	294
2	20	19	6
3	9	3	-
4	4	3	-
5	-	2	-

5. CONCLUSIONS

The results indicate that EarLyzer can form the basis for a robust ASR front-end. The experimental results show that EarLyzer can be more effective than even MFCC with dynamic features included; this can be attributed to poorer time resolution of MFCC which does not help the differential features. EarLyzer was also evaluated using a new approach to robust speech recognition based on the matched filtering formulation and the results are encouraging [4].

ACKNOWLEDGMENT

This research work is supported by a project sponsored by M/s Ericsson Inc., USA, who also provided the noisy speech database. We wish to thank M/s Ericsson for their support. We also thank our colleague Sunil, S. for providing the code for DTW and the many interesting and useful discussions we had during the course of this work.

REFERENCES

[1] Hermansky. H, "Should Recognizers have Ears?" *Speech Communication*, Vol. 25, pp. 3-27, 1998.

[2] Rabiner, L.R. and Juang, B.H. *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[3] Lockwood. P, and Boudy. J. " Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection for Robust Speech Recognition in cars", *Speech Communication, Vol.* 11, pp 215-228, 1992.

[4] Avadhanulu, J.V., and Sreenivas, T.V. "Matched Filtering Aprraoch to Robust Speech Recognition"; Accepted for Conf. Signal Proc. Commn. (SPCOM-99), Bangalore, India, Jul 1999.