Auditory Based Feature Vectors for Speech Recognition Systems

WALEED H. ABDULLA
Electrical & Electronic Engineering Department
The University of Auckland
20 Symonds Street, Auckland
NEW ZEALAND

Abstract: - Signal processing front end for extracting the feature set is an important stage in any speech recognition system. The optimum feature set is still not yet decided though the vast efforts of researchers. There are many types of features, which are derived differently and have good impact on the recognition rate. This paper presents one more successful technique to extract the feature set from a speech signal, which can be used in speech recognition systems. Our technique based on the human auditory system characteristics. It relies on the Gammatone filterbank to emulate the cochlea frequency resolution. Compared to the standard mel frequency cepstral coefficients and the perceptual linear prediction analysis front ends, our auditory based front-end yielded higher recognition rate when embedded in a continuous hidden Markov model based automatic speech recognition (ASR) system. Also our feature set has better classification characteristics when considering the F-ratio as a figure of merit.

Key-Words: - Speech Recognition, Auditory Modelling, Feature Extraction.

1. Introduction

A major problem in speech recognition system is the decision of the suitable feature set which can faithfully describe in an abstract way the original highly redundant speech signal. Several techniques have been developed so far for solving this problem. It has been known that, the cochlea, the main component of the inner ear, performs the filterbank based frequency analysis on the speech signal to extract the relevant features. Thus, most techniques are pivoting around the filterbank methodology in extracting the features. The difference in the design of the filterbank offers the extraction of different features from the signal. The main parameters in the filterbank filter design are the frequency response, which defines the shape of the filters, the centre frequency and the bandwidth. These parameters can be selected based on the human auditory system. Dominant speech analysis techniques for ASR, namely Mel frequency cepstrum [1] and perceptual linear predictive (PLP) [2], try to emulate the human auditory perception. The Mel cepstrum technique uses filters with centre frequencies spaced equally on a linear scale from 100 to 1000Hz and equally on logarithmic scale above that. Above 1000Hz each centre frequency is 1.1 times the centre frequency of the previous filter. The shape of the magnitude frequency response of each filter is normally considered triangular. The Q factor, the ratio of the centre frequency to the filter bandwidth, is constant along the whole spectral band. Each vector of log energy calculated from the filterbank outputs is processed by an inverse cosine transform to create what is called Mel frequency cepstral coefficients (MFCCs). The cosine basis attempt to approximate Karhunen-Loeve basis, which provides the necessary decorrelation between the feature vectors, and project the spectrum on directions of maximum global variability [3]. These MFCCs are considered as the extracted features from the speech signal, which in turn are presented to the speech recogniser for classification task. To improve the speech recognition rate the feature vectors are normally augmented by vectors representing the delta (speed) and the delta-delta (acceleration) of the spectral components, MFCCs [4]. The PLP differs from the Mel cepstrum in the type of filter shapes and smoothing of the short-term spectrum coming out of its filterbank. The Mel cepstrum technique uses the truncation of the MFCCs for smoothing while the PLP approximates the cubic-root compressed modified spectrum by an autoregressive model and computes the model's cepstral coefficients.

In this paper we developed a compromise between different front ends to design a model that is more

coherent to the auditory models and having the advantages of the Mel cepstrum and the PLP front ends in being fast. Our technique based on the Gammatone auditory filter bank in extracting the relevant features. Gammatone filter modelling is a physiologically based strategy followed in mimicking the structure of the peripheral auditory processing stage. It models the cochlea by a bank of overlapping bandpass filters. The performance of this technique will be measured using the F-ratio as a figure of merit to show the classification ability of our technique against the classical Mel-cepstrum and the PLP techniques. Also the recognition rate of a CHMM based ASR system will be compared by using the above three techniques. Our Gammatone based method outperformed the classical Mel-cepstrum and the PLP methods in both classification and recognition rate tests.

2. Gammatone Auditory Filterbank

Gammatone filter (GTF) modeling is a physiologically motivated strategy followed in mimicking the structure of the peripheral auditory processing stage. It models the cochlea by a bank of overlapping bandpass filters. The impulse response of each filter follows the Gammatone function shape. This function was introduced by Aertsen and Johannesma [5]. It has the following classical form:

$$h(t) = \gamma(n,b) t^{n-1} e^{-bt} \cos(wt + \phi) u(t) \qquad (1)$$

Where $\gamma(n,b)$ is a normalization constant depending on the order, n, and the bandwidth related factor, b, w is the radian center frequency, ϕ is the phase shift and $u(t)$ is a unit step function. This function has also been modified for the sake of computation simplification by removing the cosine term from the classical form [6; 7]. The name of this filter derived from its relation to the Gamma function, which has for n>0 the form:

$$\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt \qquad (2)$$

The tone is referring to the cosine term, which represents a tone at the center frequency.

Gammatone filter is very similar to the rounded exponential function, reox(p). Reox function is normally used in representing the magnitude response of the human auditory filters [8]. This function is a parameterisation form of the auditory filters response using notched noise masker technique. It is known that 3rd – 5th order Gammatone filter gives very good approximation to the reox(p) filter over a 60dB range [9]. The main advantage of Gammatone filter over the reox filter is that the former belongs to linear time invariant system family, while the later is not. This means that the Gammatone filters can be represented by a transfer function and consequently be electronically implemented while the reox filter lack this property due to unknown phase response [10]. Also, the Gammatone filterbank can very well model the non-linear frequency characteristics of the cochlea even it is belonging to the linear system family. The Gammatone function corresponding to a cochlea filter centered at 1000Hz and with bandwidth of 125Hz is shown in Fig. 1. This figure shows also that the Gammatone function is a good fit to the impulse response of the auditory nerve fibre as measured with reversed-correlation (revcor) technique [11].

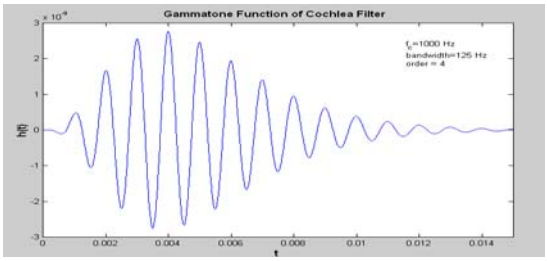

Figure 1: The Gammatone function

3. Bandwidth and centre frequency of the GTF

The bandwidth of each filter in the Gammatone filterbank is determined according to the auditory critical band (CB) corresponding to its centre frequency. The CB is the bandwidth of the human auditory filter at different characteristic frequencies along the cochlea path. The first determination of the CB was done by Fletcher in 1938 [12]. He assumed that the auditory filters were rectangular which was greatly simplifying the formulation of the signal and the noise powers within the CB. Although the rectangular critical band concept is not realistic but it is very useful. The bandwidth of the actual auditory filters can be related to it by suggesting an equivalent rectangular bandwidth (ERB) filter that has a unit height and a bandwidth *ERB*. It passes the same power as the real filter do when subjected to white noise input.

This definition of *ERB* implies the mathematical formula:

$$ERB = \int_0^\infty |H(f)|^2 \, df \qquad (3)$$

Where the maximum value of the filter transfer function, |H(f)| , is unity. Several physiologically motivated formulas have been derived for the ERB values and our preference is with that suggested by Glasberg et al. [9; 13]. It follows the following formula at centre frequency F_c:

$$ERB = 24.7(1 + 4.37 f_c) \qquad (4)$$

This formula gives the highest selectivity factor, Q factor, among all the other suggested ones. The Q factor is the ratio between the centre frequency and the bandwidth of each filter.

Thus to determine the bandwidth of each filter, which is now represented by the ERB value, the centre frequency of each filter has to be ready beforehand. In the human auditory system, there are around 3000 inner hair cells along the 3.5-cm spiral path cochlea. Each hair cell could resonate to a certain frequency within a suitable critical bandwidth. This means that there are approximately 3000 bandpass filters in the human auditory system. This resolution of filters can not be implemented practically using computational modelling techniques. However we can approximate this high resolution into some possibly implemented one. This can be achieved by specifying certain overlapping between the contiguous filters. The percentage-overlapping factor, v, will specify the number of channels, filters, required to cover the useful frequency band. This band is decided according to the requirements of the application. In our speech recognition system this band is in the range of 100 - 11025Hz, as this is the useful information distribution band. If we depend on Glassberg and Moore [13] recommendation and if we suppose that the information carrying band is bounded by f_H Hz and f_L Hz with v overlapping spacing the number of filters will be:

$$N = \frac{9.26}{v} \ln \frac{f_H + 228.7}{f_L + 228.7} \qquad (5)$$

Then the center frequency can be calculated by

$$f_c = -228.7 + (f_H + 228.7) e^{-\frac{vn}{9.26}} \qquad (6)$$

where $1 \le n \le N$

Having decided the locations of the center frequency of each filter the bandwidth can be calculated from (4) and we can now proceed to the implementation stage.

4. Gammatone Filter Implementation

The previous sections described a physiologically motivated way for deciding the bandwidth and the center frequency of each filter in the Gammatone filterbank. The implantation of a band pass filter from its time domain function is a straightforward procedure in signal processing. It is simply started by finding the Laplace transform of the Gammatone function then map it into the digital form using bilinear transform or impulse invariant transform. There are several methods in representing and implementing the Gammatone function. Lyon suggested an all pole version by discarding the zeros from the transfer function of the Gammatone filter aiming for simple parameterization [10]. Lyon all pole version reduces the computation but on the cost of loosing selectivity sharpness at low frequency. Other form was also suggested by Cooke, [7], in which he used the complex data to realize fourth order filter. The computations still as that of the eighth order filter with real data. Cooks method needs pre-multiplication of the input signal by a complex exponential at the specified center frequency, filtering with a beseband Gammatone filter, post-multiplication by that exponential. Malcolm Slaney describes one simple way of implementation procedure of the Gammatone based filters [14]. Fourth order Gammatone filter is also used in the design as it gives the best reox function fit. It requires eighth order digital filter to realize. Our preference is with Slaney method because it preserves the original form of the Gammatone filter, and of the simplicity of implementation. The frequency response of a 20-channel filterbank, covering 100-11025 Hz band, after the pre-emphasis by the equal loudness curve is shown in Fig. 2.

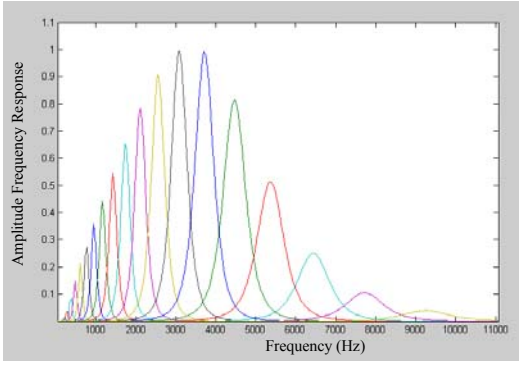

Figure 2: Frequency response of a Gammatone filterbank

The bandwidth of the channels is logarithmically proportional with the centre frequency.
Fig. 3 shows the relation between the channel number, the centre frequency and the bandwidth. The tips of the filters moving along the centre of the horny shaped curve. The highest the channel number is the lowest the centre frequency and bandwidth are, which is in consistency with equation (6). The bandwidths of the filters are logarithmically varied with the channel numbers.

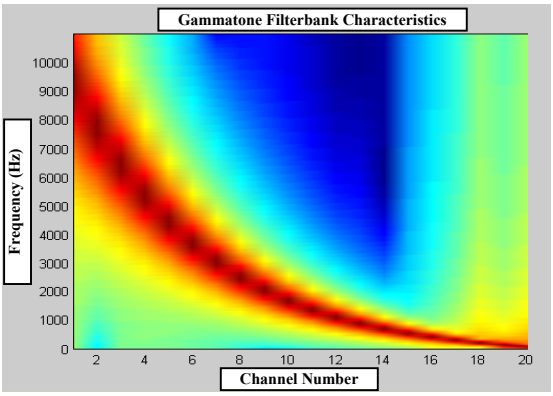

Figure 3: Gammatone filterbank characteristics

5. Speech analysis using Gammatone filterbank

The physiologically motivated Gammatone filters can be used as weighting coefficients for speech signals. In this case the energy within each filter is calculated by finding the magnitude of the Fourier transform of the speech signal and multiplying it by its corresponding weighting filter. The filters' outputs are subjected to equal loudness pre-emphasis filter. From this stage we experimented two options. The first option, Gamma-cepst is rapping the energy spectra into cepstral coefficients domain using the inverse cosine transform. This transformation produces highly uncorrelated features, which are necessary for the HMM processing. To reduce the high dimensionality of the analyzed speech into low dimensionality space, a smoothing by truncating the output coefficients to 13 coefficients is necessary. The second option, Gamma-PLP, is to augment similar steps used in preparing the PLP coefficients. The block diagrams of both options are depicted in Fig. 4.

6. Evaluation based on the F-ratio

The F-ration is a measure that can be used to evaluate the effectiveness of a particular feature. It has been widely used as a figure of merit for feature selection in speaker recognition applications [15]. It is defined as the ratio of the between-class variance (B) and the within-class variance (W).

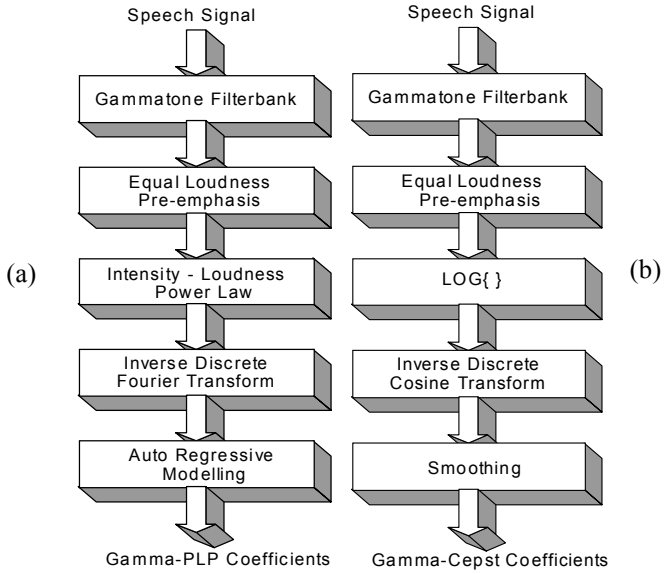

Figure 4: Block diagrams of two feature extraction paradigms. (a) Gamma-PLP (b) Gamma-cepst

In the contest of feature selection for pattern classification, the F-ratio can be considered as a strong catalyst to select the features that maximise the separation between different classes and minimise the scatter within these classes. The F-ratio technique can be formulated as follows:
Let us consider that the number of training feature vectors, training patterns, in the j^{th} class of K classes be equal to N_j. Thus the F-ration of the i^{th} feature can be defined by

$$F_i = \frac{B_i}{W_i} \qquad (7)$$

where B_i is the between-class variance and W_i is the pooled within-class variance of the i^{th} feature. These can be mathematically defined by

$$B_i = \frac{1}{K}\sum_{j=1}^{K}(\mu_{ij} - \mu_i)^2 \qquad (8)$$

$$W_i = \frac{1}{K}\sum_{j=1}^{K}W_{ij} \qquad (9)$$

where μ_{ij} and W_{ij} are the mean and variance of the i^{th} feature, respectively, for the j^{th} class, and μ_i is the overall mean of the i^{th} feature.

In our approach in using the F-ratio we make use of the HMM properties to facilitate the implementation of this technique. The HMM technique used is implicitly considering the Gaussian behaviour of the feature vectors which satisfies one condition needed by the F-ratio method. The second condition, uncorrelation, is satisfied by using the diagonal covariance within the structure of the HMM.

In our approach we applied the F-ratio, formula (7), on each model, corresponding to a certain word, considering each state as a separate class. Then we averaged the resultant F-ratios of the different words. In this case K refers to the number of states in the HMM. The F-ratio averaging is straightforward according to the formula

$$F^{ave} = \frac{1}{H}\sum_{i=1}^{H} F_i \qquad (10)$$

where H is the number of models to be dealt with.

The averaged F-ratio values can be sorted into descending order then the top Q features are selected, which simply determine the most vital features within the whole set of features. The number of coefficients of the full feature vector is Q=39, in proportion of 13 (power and 12 MFCCs) with their delta and delta-delta coefficients. The F-ratio of 10 models and their average is depicted in Fig. 5.

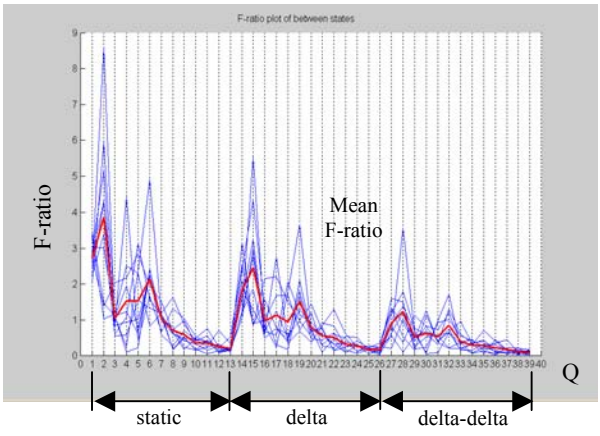

Figure 5: F-ratio of the between states procedure. The thick red line indicates the mean of the between states F-ratio.

It is obvious from this figure that the static coefficients set, Q_1 to Q_{13}, are more important than the dynamic delta coefficients set, Q_{14} to Q_{26}, which in turn more important than the delta-delta coefficients set , Q_{27} to Q_{39}. This motivated us to select the most prominent coefficients from the feature vectors. We selected the top 28 ranked coefficients in proportion of staic=11, delt=9, and delta-delta=8. This selection has proved to be better than the original 39 coefficients and be used in all our experiments.

We compared the F-ratio characteristics of the Mel-cepst, Gamma-cepst, Gamma-PLP, and PLP models to evaluate the classification performance of them. Fig. 6 shows this comparison which indicates that the models performance from the highest to the lowest is in the following order: Gamma-cepst, Gamma-PLP, PLP, and Mel-cepst. Their corresponding F-ratio total means, mean(F^{ave}), are 1.57, 1.45, 1.30, and 1.19.

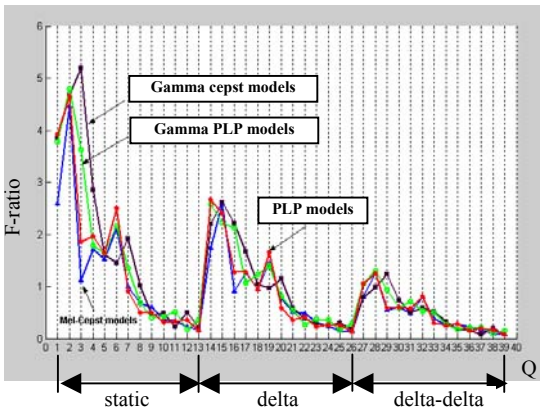

Figure 6: Classification properties based on F-ratio calculations of different feature extraction paradigms.

To consolidate the consistency of classification property with the recognition performance we embedded the above four features into a standard CHMM based ASR system. The models are left-to-right, 9 states, 5 mixture models suitable for medium size, speaker independent isolated word recognition [16;17].

The testing datasets are DATASET-I that includes the 35 English digits, and DATASET-II that includes 105 randomly selected words. The recognition rates are depicted in Table 1.

	DATASET-I	DATASET-II
Mel-cespt	97.3	92.6
PLP	98.1	92.9
Gamma-PLP	98.8	95.7
Gamma-cepst	99.6	98.2

Table 1: Recognition rate performance results of the four feature sets.

For DATASET-I all the paradigms have almost the same performance while in DATASET-II, larger

size data, the Gamma-cepst outperforms the features, which is consistent with the F-ratio results.

7. Conclusion

An auditory motivated technique has been described to extract significant feature sets from the speech signal. It is mainly based on the Gammatone filterbank. Gammatone Auditory filter banks are non-uniform bandpass filters, designed to imitate the frequency resolution of human hearing. Two paradigms shown in Fig. 4 have been implemented and tested. They outperform their classical counterparts, i.e. mel frequency and PLP techniques. The classification performances have been tested since they are strong cue to the recognition performances. Intuitively, the more distant the classes are from each other, the better the chance of successful recognition of class membership of patterns. It is reasonable, therefore, to select as the feature space that d-dimensional subspace of the pattern representation space in which the classes are maximally separated.

In comparison to the conventional mel frequency and PLP techniques, Gammatone based features were embedded in a standard CHMM based ASR system and the recognition rates were calculated. Table 1 shows that our technique outperforms the conventional feature based ASR systems and the Gamma-cepst features are the best performing paradigm.

The F-ratio computation has two roles the first one is to show the classification performances. The second one is to select the most prominent features. We have seen that using 28 coefficients in proportion of: static=11, delta=9, and delta-delat=8 are performing better than the original 39 coefficients.

References:

[1] Davis, B., and Mermelstein, P. (1980). "*Comparison of parametric representations for monosylabic word recognirion in continuously spoken sentences.*" IEEE Trans. ASSP, 28(4), 357-366.

[2] Hermansky, H. (1990). "*Perceptual linear predictive (PLP) analysis for speech.*" J. Acoust. Soc. Am., 87, 1738-1752.

[3] Hermansky, H., and Malayath, N. (1998). "*Spectral basis functions from discriminant analysis.*" ICSLP'98, Sydney, Australia.

[4] Furui, S. (1986). "*Speaker-independent isolated word recognition using dynamic features of speech spectrum.*" IEEE Trans. ASSP, 34, 52-59.

[5] Aertsen, A., and Johannesma, P. (1980). "*Spectro-temporal receptive fields of auditory neurons in the grass frog. I. Characterization of tonal and natural stimuli.*" Biol. Cybern., 38, 223 - 234.

[6] Solbach, L. (1998). "*An architecture for robust partial tracking and onset localization in single channel audio signal mixes,*" PhD thesis, Technical University of Hamburg-Harburg, Germany.

[7] Cooke, M. (1993). "*Modelling Auditory Processing and Organization"*, Cambridge University Press, U.K.

[8] Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (1982). "*The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram and speech threshold.*" J. Acoust. Soc. Am., 72, 1788 - 1803.

[9] Patterson, R. D. (1994). "*The sound of a sinusoid: Spectral models.*" J. Acoust. Soc. Am., 96(3), 1409 - 1418.

[10] Lyon, R. F. (1997). "*All-pole models of auditory filtering.*" Diversity in auditory mechanics, Lewis, ed., World Scientific Publishing, Singapore, 205 - 211.

[11] de-Boer, E., and H. R. de Jongh. (1978). "*On cochlea encoding: Potentialities and limitations of the reverse-correlation technique.*" J. Acoust. Soc. Am., 63(1), 115 - 135.

[12] Allen, J. B. (1995). "*Speech and hearing in communication"*, ASA edition, Acoustical Society of America, New York.

[13] Glasberg, B. R., and Moore, B. C. (1990). "*Derivation of auditory filter shapes from notched-noise data.*" Hearing Research, 47, 103 - 108.

[14] Slaney, M. (1993). "*An efficient implementation of the Patterson-Holdsworth auditory filter bank.*" *Apple Computer Technical Report #35*, Apple Computer Inc.

[15] Paliwal, K. K. (1992). "*Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer.*" Digital Signal Processing, 2, 157-173.

[16] Abdulla, W.H. and N.K. Kasabov (1999). *"Two pass hidden Markov model for speech recognition systems".* Proc. ICICS'99. Singapore, Paper #175.

[17] Abdulla, W.H., (2002) *"Signal Processing and Acoustic Modelling of Speech Signal for Speech Recognition Systems"*, PhD Thesis, Information Science Department, University of Otago, New Zealand.