SPEECH RECOGNITION USING FILTER-BANK FEATURES

Sourabh Ravindran, Cenk Demiroglu, and David V. Anderson

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta GA 30332

ABSTRACT

Mel-frequency cepstral coefficients (MFCC) have been shown to be very useful in tasks of speech recognition and are the preferred features in state of the art speech recognition systems. We present features derived from filter bank outputs whose performance is comparable to that of MFCCs for connected digit recognition using a Hidden Markov Model (HMM) based speech recognition system. The feature extraction method we present is easily implementable in floating gate analog VLSI circuitry which makes it a viable option for low power speech recognition tasks.

1. BACKGROUND

Filter bank energies (FBEs) were quite popular in the early days of speech recognition [1] but over the last couple of decades MFCCs have emerged as the most popular features for speech recognition. In addition to the performance gain, one of the main reasons for discarding FBEs in favor of other features has been the computational gain achieved from the efficient implementation of these features. The recognition models used in the early days were not Gaussian-based HMM speech recognition systems and perhaps for this reason the filter bank outputs were never decorrelated. Now with advances in speech recognition back-ends and the advent of floating-gate analog VLSI technology it might be worth revisiting the FBEs. The implementation of FBEs in analog VLSI technology would yield a huge gain in terms of power consumption. Further, Searle *et al.* [2] have shown that using filter-banks followed by a peak detector can capture voiced onsets and curvature of energy peaks (formant tracks) that might be very useful for speech recognition.

Linear prediction coefficient (LPC) features were also quite popular. White and Neely [1] compared FBEs with LPC for isolated word recognition . They used one-third octave filters with 20 frequency channels. Output of each channel was energy smoothed, noise subtracted and subjected to log amplitude scaling. For the control case, 14 LPC coefficients were calculated using autocorrelation. Dynamic programming was used for time-alignment. It was reported that both LPC and filter-bank methods gave similar performance. However, the recognition time per utterance was lesser for LPC features and the data rate was approximately 3 times lower than that of filter-bank features. Presumably these advantages led to FBEs being replaced by LPCs.

Dautrich *et al.* [3] studied different filter design choices (number of channels, type of filter, filter spacing etc.) on the performance of the recognizer. Eight uniform and five non-uniform filters with varying amounts of overlap were considered. It was reported that a 15 channel uniform filter-bank and a higly-overlapping 13 channel non-uniform critical band filter-bank performed best on a 39 word alphadigit vocabulary. It was also reported that an 8th order LPC based recognizer performed better than the filter-bank based approach. However, it is interesting to note that the performance of LPC based recognizer deteriorated faster than that of filter-bank features at low SNRs.

More recently, Nadeu *et al.* [4], compared decorrelated FBEs with MFCCs. The decorrelated FBE are obtained by filtering the logarithm of the filter bank (discrete fourier transform) outputs. The filter (1^{st} order highpass filter) used provides both variance equalization and decorrelation. Continuous observation density HMM was used for recognition. For the control case, 8 MFCCs were used. The authors claim that for their system 8 MFCCs were found to be optimum. It was reported that decorrelated FBEs performed on par with MFCCs.

Paliwal [5] also compared decorrelated FBE with MFCCs. A linear predictor was used for decorrelation of the FBEs and an FIR (highpass) filter was used for liftering the FBE features. The effect of using different lifters was studied. It was reported that FBEs performed as well as MFCCs in clean and noisy conditions.

It is fair to conclude from the referenced studies that FBEs do offer some advantages over MFCCs from a purely signal processing point of view. It appears that FBEs are more robust to noise. Filtering of FBEs allows for implicit weighting of features in cepstral domain that might be useful for speech recognition. The rest of the paper is organized as follows, section 2 introduces the feature extraction pro-

0-7803-8104-1/03/$17.00 ©2003 IEEE

cess and section 3 explains the silicon implementation of the feature extraction process. Section 4 describes the experiments followed by results and conclusion.

2. FEATURE EXTRACTION

The feature extraction process explained herein is loosely based on a model of the human auditory system described by Yang *et al.* [6]. The bandpass filter bank approximates the response of the basilar membrane. The velocity coupling of the cochlear fluid with the cilia of the inner hair cell is modeled by a time derivative. We contend that this time derivative can be incorporated into the filter bank by having a 40 dB/decade roll-off as opposed to a 20 dB/decade roll-off usually used to model the basilar membrane response. The non-linearity of the neurons in the cochlear nucleus is modeled by a half-wave rectifier and the limited response of these neurons to fast temporal fluctuations is modeled by a lowpass filter. We combine these two stages into a peak detector stage. The lateral inhibition described in [6] has been ignored. We introduce a logarithmic compression to incorporate the effect of the outer hair cells.

Figure 1 shows the block diagram of the feature extraction process. The speech signal is passed through a bank of bandpass filters. There are 16 channels with one-third octave spacing covering the frequency range from 100 Hz to 4000 Hz. The output of each channel is passed through a peak detector and logarithm of the peak detector output is computed. The features are obtained by multiplying the output with the eigenvectors of the autocorrelation matrix of the speech input. The autocorrelation matrix is computed by taking the average over the entire training data set. A singular value decomposition is performed to obtain the eigenvector matrix. For the purpose of comparison discrete cosine transform (DCT) was also used to perform the decorrelation.

3. SILICON IMPLEMENTATION

One of the biggest advantages of this approach is that it can be implemented using programmable floating–gate analog VLSI circuitry [7]. The analog structure provides parallel computation ability and consumes orders of magnitude lesser power than a similar digital implementation. Since the decorrelation matrix is computed offline we can store these values on the floating gate nodes of the analog circuitry to obtain real time performance in hardware. The circuit can be interfaced with digital systems and the features can directly be used for digital speech recognition.

Capacitively Coupled Current Conveyer Second–Order Section (C^4 SOS) can be used to build the bandpass filterbank. The C^4 SOS is a continuous–time bandpass filter with electronically tunable corner frequencies and ± 40

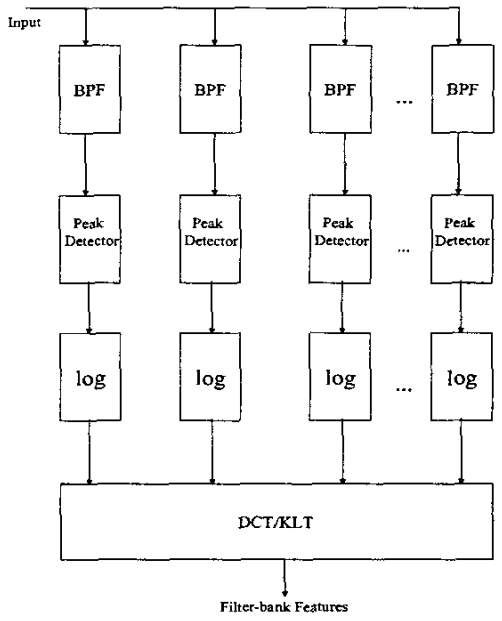

Fig. 1. Feature extraction.

dB/decade or greater roll–offs. The corner frequencies can be set independently of each other, and the bandwidth can therefore be tuned at will. Each corner can have its own Q peak, or, if the corners are brought close together, a very tight bandwidth with a single Q peak develops. This leads to further isolation of any given frequency. An array of these C^4 SOS's with exponentially spaced center frequencies forms a good model of the frequency response of the human cochlea where signals are decomposed with filtering processes that have $Q \approx 30$ [8]. A C^4 SOS is as shown in fig. 2.

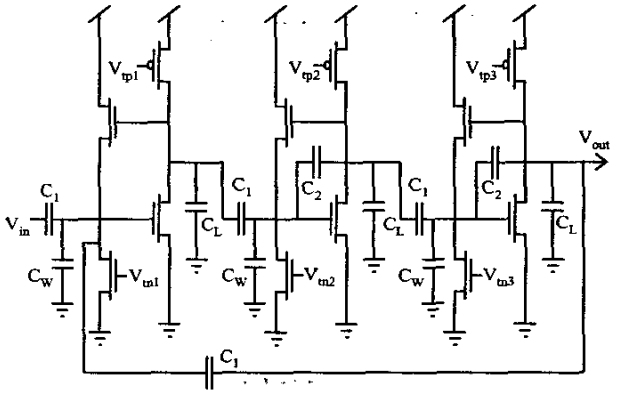

Fig. 2. C^4 second order sections.

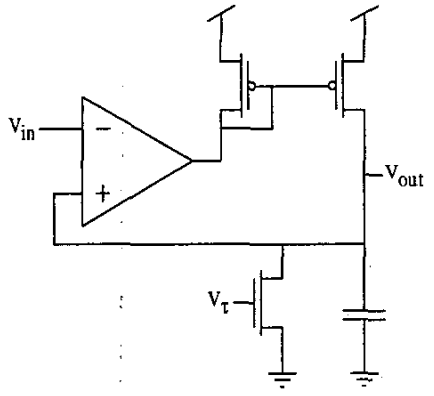

Fig. 3. Analog peak detector circuit.

The peak detector circuit is as shown in fig. 3. The time constant of the peak detector can be adjusted for each of the channels. The time constant can be varied by adjusting V_τ. Figure 4 shows the output from the a selected channel of the analog bandpass filter-bank and the peak detector circuits for a speech input.

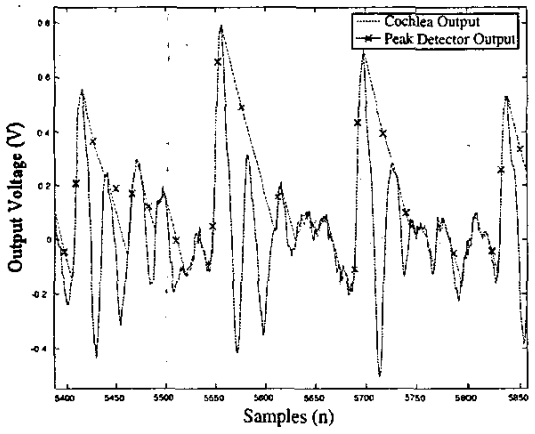

Fig. 4. The output from the analog bandpass filter-bank (cochlea) and the peak detector circuits are shown for a speech input.

4. EXPERIMENT

The filter-bank features were compared with MFCCs for a connected digits recognition task using the TIDIGITS database. The ISIP speech recognition engine was used for the experiment. The filter-bank features had to be filtered with a median filter to obtain 16 features every 10 msecs, this was done to obtain a common comparison framework with MFCCs. For the recognition, word based templates were used with 16 mixtures per state.

5. RESULTS AND CONCLUSION

We have found that while FBEs are not as good as MFCCs, their performance is comparable to that of MFCCs. The recognition results are as shown in Table 1. Using DCT instead of KLT leads to a slightly lower recognition rate of 98.2%. However in defence of FBEs it must be mentioned that smoothing the filter-bank output might be hurting the recognition results. In our current approach the temporal information available from the envelope-based method is not being exploited. Also unlike in [4] and [5] we do not lifter the features. Implicit weighting of the features in the cepstral domain through filtering in the frequency domain might further help improve the recognition rate. The results obtained herein merits further study into the use of FBEs for speech recognition. The temporal information provided by our approach could be used to detect syllable onsets and offsets which have been shown to be useful for speech recognition [9]. In the current recognition systems the syllabification of lexical items is derived from a more abstract phonological representation, this problem could be addressed using the feature extraction method described here.

Recognition Rates	
MFCC	Filter-bank Features
99.8 %	98.6 %

Table 1. Table showing the recognition rates for MFCCs and filter-bank features.

6. REFERENCES

[1] G. White and R. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," in *International Conference on Acoutics, Speech and Signal Processing*, Apr 1976, vol. 24, pp. 183–188.

[2] C. Searle, J. Jacobson, and S. Rayment, "A phoneme recognition system based on human audition," in *International Conference on Acoutics, Speech and Signal Processing*, Apr 1978, vol. 3, pp. 557–560.

[3] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the use of filter bank features for isolated word recognition," in *International Conference on Acoutics, Speech and Signal Processing*, May 1983, vol. 8, pp. 1061–1064.

[4] C. Nadeu, , J. Hernando, and M. Gorricho, "On decorrelation of filter-bank energies in speech recognition," in *Proceedings of Eurospeech'95*, Sept 1995, vol. 1, pp. 1381–1384.

[5] K. K. Paliwal, "On the use of filter-bank energies as features for robust speech recognition," in *International Symposium on Signal Processing and its Applications*, Brisbane, Australia, Aug 1999, vol. 2, pp. 641–644.

[6] Xiaowei Yang, Kuansan Wang, and Shihab Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[7] P. Smith, M. Kucic, R. Ellis, D. Graham, and P. Hasler, "Mel-frequency cepstrum encoding in analog floating-gate circuitry," in *International Symposium on Circuits and Systems*, Phoenix, AZ, May 2002, vol. 4, pp. 671–674.

[8] David Graham and Paul Hasler, "Capacitively-coupled current conveyer second-order section for continuous-time bandpass filtering and cochlea modeling," in *International Symposium on Circuits and Systems*, May 2002, vol. 5, pp. 485–488.

[9] Su-Lin Wu, M.L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *International Conference on Acoutics, Speech and Signal Processing*, Apr 1997, vol. 2, pp. 987–990.