ROBUST WORD RECOGNITION USING THREADED SPECTRAL PEAKS

B. Strope and A. Alwan

Electrical Engineering Department, UCLA, 405 Hilgard Ave., Los Angeles CA, 90095
bps@ucla.edu, alwan@icsl.ucla.edu

ABSTRACT

A novel technique which characterizes the position and motion of dominant spectral peaks in speech, significantly reduces the error-rate of an HMM-based word-recognition system. The technique includes approximate auditory filtering, temporal adaptation, identification of local spectral peaks in each frame, grouping of neighboring peaks into threads, estimation of frequency derivatives, and slowly updating approximations of the threads and their derivatives. This processing provides a frame-based speech representation which is both dependent on perceptually salient aspects of the frame's immediate context, and well-suited to segmentally-stationary statistical characterization. In noise, the representation reduces the error-rate obtained with standard Mel-filter-based feature vectors by as much as a factor of 4, and provides improvements over other common feature-vector manipulations.

1. INTRODUCTION

The eigenfunctions of a resonating vocal tract are manifested acoustically as formants in speech. The analysis of formants has provided significant insights into speech production mechanisms, and motivation for speech coding algorithms. Referring to automatic speech recognition (ASR) in 1981, D. Klatt wrote [1]:

> "These schemes will succeed only to the extent that metrics can be found that are (1) sensitive to phonetically relevant spectral differences such as those caused by formant frequency changes, and (2) relatively insensitive to phonetically irrelevant spectral differences associated with a change of speaker identity or recording conditions."

Although various compensation schemes for changing acoustic environments are often used, the predominant characterization of speech for statistical speech recognition is based on sequences of short-time (10-20 ms) spectral estimations, which characterize the coarse spectral envelope of each successive frame [2]. This representation is only an implicit characterization of the formant structure of speech, and as such, does not provide direct access to the phonetically relevant formant motion described above. Explicit characterizations of speech dynamics typically focus on the motion of the cepstral representation of the short-time spectral estimates [3,4], and thereby parameterize changes in the 'complete' spectral shape and not the specific (potentially robust) formant motion.

More direct formant tracking typically involves first identifying local spectral peaks in a sequence of spectral estimations [5,6]. Alternatively, Teager energy operators [7-10], Hilbert Transforms and Wigner Distributions [11], as well as changes in the cross-correlation of the temporal fine-structure between neighboring auditory frequency channels [12] have been used to identify formant frequencies in speech. Formant tracks are then pieced together using heuristics [5,6], hidden Markov models [13], or the minimization of a cost function [14]. The two-stage process has also been collapsed to one using extended Kalman filters [15,16]. Unfortunately formant tracking systems are often non-robust; they are only occasionally evaluated in noise, and are rarely tested in the context of an ASR task.

Processing schemes which enhance the representation of spectral dynamics and, more specifically, changing spectral peaks (e.g. [17,20]) have been proposed. While such sensitivity may be phonetically relevant, the characterization of the formant motion is usually implicit. Formant motion is only weakly characterized by the temporal derivative of the overall spectral estimate, and by the sequence of underlying states in the statistical model. Neither of these is a direct characterization, and neither provides an obvious means to exploit the dominant frame-to-frame correlations of local spectral peaks. Finally, context dependent spectral representations may, in general, be poor matches to ASR algorithms which rely on the characterization of segmentally stationary statistics. A more direct parameterization of the motion of spectral peaks, on the other hand, may prove to be a better match.

The algorithm described here introduces a simple and robust form of formant tracking, and augments the frame-based feature vector used for ASR with an explicit parameterization of the formant position and motion.

2. ALGORITHM OVERVIEW

A block diagram of the processing stages in the algorithm is shown in Fig. 1. The initial filtering and subsequent liftering process the wide-bandwidth speech signal at the full sampling rate (12k samples/sec). All remaining processing occurs at the down-sampled frame rate (100 frames/sec) with substantially lower computational complexity.

2.1 Filtering and Adaptation

The filtering stage, after [18], is implemented by integrating

0-7803-4428-6/98 $10.00

power spectrum estimates weighted by triangular filters that have bandwidths of 100 Hz for center frequencies below 1 kHz, and bandwidths of 0.1 times the center frequency above 1 kHz. The resulting frequency resolution is therefore linear below 1 kHz, and logarithmic above 1 kHz

The adaptation stage for each frequency channel acts as an automatic gain control which incrementally adjusts an additive logarithmic offset to reduce the distance to a target input/output point. Adaptation emphasizes onsets and represents changing spectral peaks more strongly than static ones. Together, these two stages significantly affect how spectral peaks are identified and processed in subsequent stages. Fig. 2.a includes a spectrogram of the four digits, "nine six one three," at 10-dB SNR, after filtering and adaptation.

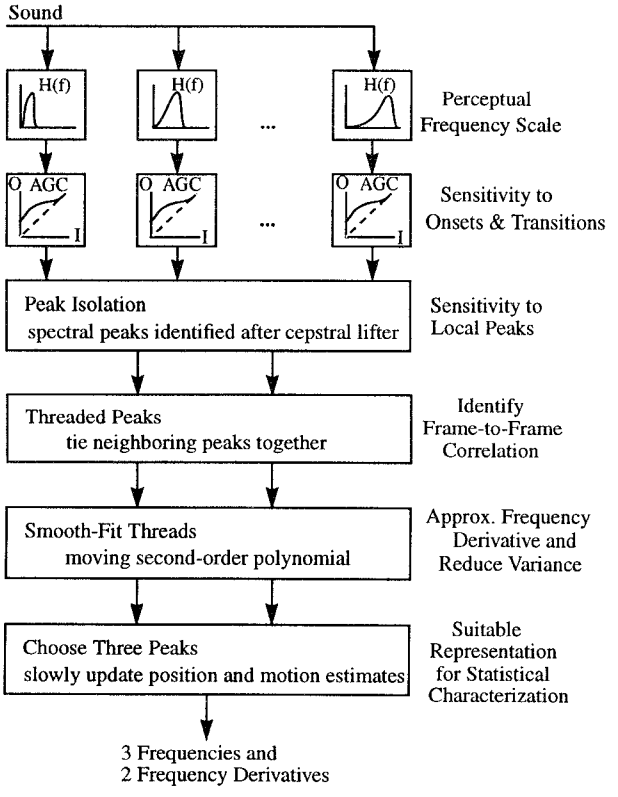

Figure 1. Processing overview.

2.2 Peak Isolation

Local spectral peaks are first identified independently in each frame by finding the local maxima in the log-spectral estimate, after raised-sin cepstral liftering [19]. For each peak (marked in Fig. 2.a), the frequency position and log magnitude are stored. Because the raised-sin cepstral lifter alters the level of the local spectral peak, the log-magnitude value is taken from the corresponding frequency position in the spectral estimate before raised-sin liftering.

In Fig. 2.a, note the relatively strong temporal correlation between the frequency positions of the local spectral peaks through formants and formant transitions.

2.3 Threading Peaks

This is the first of two stages which group peaks based on their spectro-temporal proximity. The task is to connect the spectral peaks together in time into *threads*, and the approach used here is a form of dynamic programming. Each peak (in each frame) is connected to the closest thread that extends into at least one of the last two frames. If the frequency distance to the closest thread is greater than approximately 10% of the total (warped) frequency range, then a new thread is started. If no peak connects to the end of a given thread for two successive frames, then that thread is ended. Fig. 2.b shows a moving seven-point second-order polynomial fit to each resulting thread. For each thread that includes at least four peaks, the temporal derivative as implied by the moving second-order polynomial is also stored.

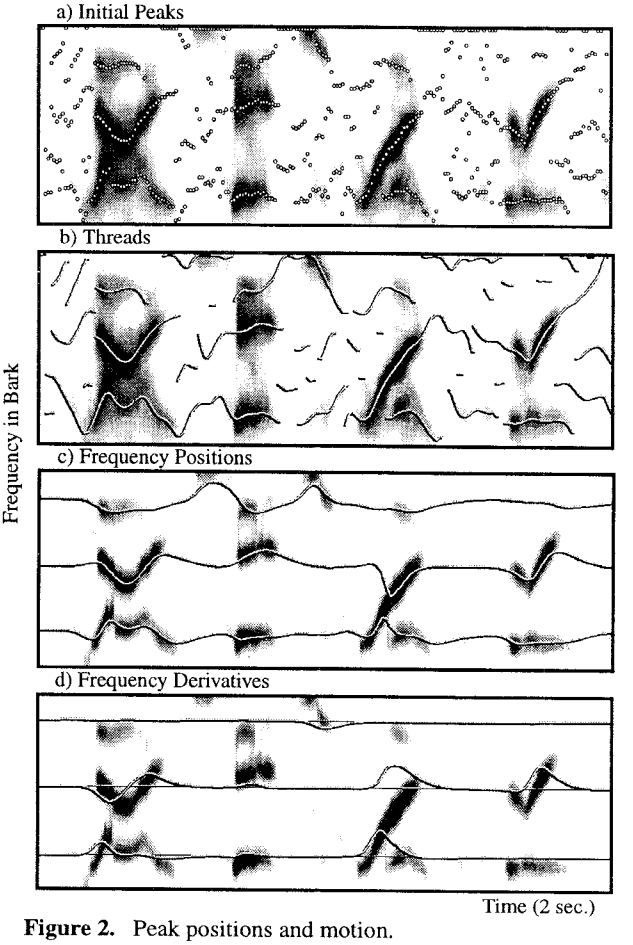

Figure 2. Peak positions and motion.

2.4 Choosing Three Peaks

The second stage imposes a structure on the threads enabling a more systematic characterization, and also attempts to reduce their variance. Threads from the first stage start and end somewhat

randomly, which makes storing them for analysis or comparison not obvious. Also, there is significant variance in the reliability of the thread measurements. That is, dominant formant transitions are tracked more reliably than small peaks in background noise.

The second stage limits the representation of the threads to three peaks in frequency for each frame. Three new *tracks*, centered at relatively low, medium, and high frequencies, are used to represent the information from the threads. The log magnitude of the original spectral peak is used when integrating frequency positions and derivatives from the corresponding thread. This introduces an inertial response that updates more quickly to more dominant peaks.

In the implementation, each track is assigned a center frequency, or DC offset. The three center frequencies are equally spaced on the warped frequency scale. At each frame the frequency position of the track incrementally adjusts toward the closest thread in that frame. The increment of adjustment is a sigmoidal function of the magnitude of the thread. The equation that describes this adjustment is:

$$f[n] = \alpha\, p[n] + (1-\alpha)\,(0.9\, f[n-1] + 0.1\, f_0),$$

where n is the frame index, $f[n]$ is the frequency of the track, $p[n]$ is the frequency of the nearest peak, f_0 is the center frequency or DC offset, and the variable α, which controls the rate of the increment, is a sigmoidal function of the log magnitude of the peak. Ignoring the DC offset, the equation describes a non-constant coefficient first-order low-pass filter. The sigmoid maps log magnitude to the appropriate (0,1) interval, so that the filter changes from low-pass to all-pass. Because the log magnitude of the peak is measured after the adaptation stages, transitions and onsets incur the most abrupt track changes.

An identical structure is used to track the frequency derivatives of the threads. For each of the three tracks, the current frequency derivative estimate is incrementally updated to the derivative measured at the closest peak. The size of the increment is a sigmoidal function of the log magnitude of the peak. A final low-pass filter with a cut-off at 15 Hz is applied both to the three frequency tracks, and to the three derivatives. Fig. 2.c shows the frequency positions for the three tracks, and Fig. 2.d shows the frequency derivatives.

This parameterization of the motion of local spectral peaks differs from more traditional formant tracking [5,6] in several ways. The initial filtering and adaptation greatly influence the resulting spectro-temporal representation. The frequency resolution is warped to a perceptual scale, and signal dynamics play a significant role in determining which peaks are identified. The two-stage process to identify the final tracks, is aimed at identifying the robust, slowly-varying information which is likely to be highly correlated with underlying articulator motion. The tracking process also includes an inertial component dependent on the magnitude of the (adapting) response of the peak. Initial frequency derivative estimates are calculated before the imposition of explicit frequency ranges, reducing the influence of artifacts from these heuristics on the derivative estimates. Finally, by limiting the representation to three peaks with centers equally-spaced on the warped frequency scale, some of the complications introduced by the merging and splitting of higher formants are avoided.

3. RECOGNITION EVALUATIONS

A discrete-word recognition task was used to evaluate the robustness of a variety of processing schemes. Digits from the TI-46 database were used at a random offset within roughly two seconds of silence. The system was, therefore, required to distinguish the speech from the background.

Each digit was modeled using a six-state left-to-right HMM with continuous Gaussian densities. The forward/backward algorithm was used to estimate feature-vector means and state transition probabilities. A diagonal covariance estimate, from the entire training set, determined a global observation variance. Models derived from both clean and noisy data were used simultaneously, with the most probable model determining the digit recognized.

For all processing schemes, the feature vector included 12 cepstral coefficients (c_0 was excluded), and 13 cepstral derivatives. Three peak frequencies, and two frequency derivatives were also included in the 'threaded' evaluations. The frequency derivative of the highest peak was excluded because it had little variance.

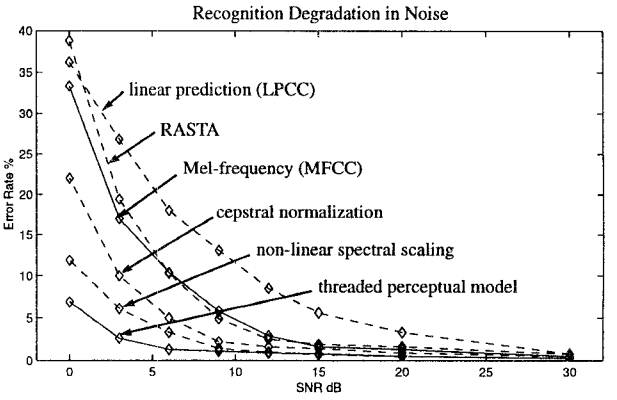

Figure 3. Recognition error rates.

Background noise, shaped to match an estimate of the long-term average speech spectrum, was added to corrupt the speech signal for these evaluations. In addition to linear prediction cepstral coefficients, Mel-frequency cepstral coefficients, and the relative-spectra (RASTA) technique [20], tests were performed using MFCC with spectral subtraction, spectral scaling, non-linear spectral scaling, cepstral mean subtraction, and cepstral normalization. The top two of these last five are included in Figure 3.

In spectral subtraction, an estimate of the background power spectrum is subtracted from each frame. Spectral scaling performs a similar subtraction using log-magnitude power-spectrum estimates. Non-linear spectral scaling was implemented by averaging

the estimate obtained after spectral scaling with another post-scaling estimate which was scaled (again) to provide the same peak log-magnitude difference from the background across all tokens. The weights used in the averaging for non-linear spectral scaling were iteratively adjusted to improve recognition performance. In cepstral mean subtraction, a long-term average cepstral vector is subtracted from each frame. Cepstral normalization was implemented by scaling the length of each cepstral vector to unity. The recognition system was completely retrained for each type of processing.

4. Conclusion

A general processing scheme is presented which may provide a more phonetically relevant parameterization of speech. The processing includes filtering, adaptation, and peak isolation, together with a relatively simple two-stage process which parameterizes the motion of dominant local spectral peaks. The algorithm has low computational complexity and is robust to background noise. The resulting characterization may provide an improved match to segmentally-stationary statistical characterization, and increases the noise-robustness of a discrete-word recognition system.

Acknowledgments

This work was supported in part by NIDCD grant DC02033, and the NSF.

References

[1] D. Klatt, "Prediction of perceived phonetic distance from short-term spectra--a first step," *J. Acoust. Soc. Am.*, vol. 70, Suppl. 1, S59, 1981.

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

[3] L. Deng, "Integrated optimization of dynamic feature parameters for hidden Markov modeling of speech," *IEEE Signal Processing Letters*, vol. 1, no. 4, pp. 66-69, April 1994.

[4] L. Deng, M. Aksmanovic, X. Sun, C. F. J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 4, pp. 507-520, Oct. 1994.

[5] R. W. Schafer, L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. of Am.*, vol. 47, no. 2.2, pp. 634-648, Feb. 1970.

[6] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. ASSP-22, no. 2, pp.135-1341, April 1974.

[7] P. Maragos, J. F. Kaiser, T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532-1550, April 1993.

[8] J. T. Foote, D. J. Mashao, H. F. Silverman, "Stop classification using DESA-1 high resolution formant tracking," *Proceedings of IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol. 2., pp. 720-723, April 1993.

[9] H. M. Hanson, P. Maragos, A. Potamianos, "A system for finding speech formants and modulations via energy separation," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 3, pp.436-443, July 1994.

[10] A. Potamianos, P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. of Am.*, vol. 99, no. 6, pp. 3795-3806, June 1996.

[11] P. Rao, "A robust method for the estimation of formant frequency modulation in speech signals," *Proceedings of IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol. 2, pp. 813-816, May 1996.

[12] L. Deng, I. Kheirallah, "Dynamic formant tracking of noisy speech using temporal analysis on outputs from a nonlinear cochlear model," *IEEE Trans. on Biomed. Engin.*, vol. 40, no. 5, pp. 456-467, May 1993.

[13] G. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. ASSP-34, no. 4, pp. 709-729, Aug. 1986.

[14] Y. Laprie, M.-O. Berger, "A new paradigm for reliable automatic formant tracking," *Proceedings of IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol. 2., pp. II.201-204, April 1994.

[15] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter." *Proceedings of IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol. 2., pp. 1229-1232, April 1986.

[16] M. Niranjan, I. Cox, S. Hingorani, "Recursive tracking of formants in speech signals," *Proceedings of IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, vol. 2., pp. II.205-208, April 1994.

[17] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 5, no. 5, pp. 451-464, Sept. 1997

[18] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. 28, pp. 357-366, Aug. 1980.

[19] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 35, pp. 947-954, July 1987.

[20] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 578-589, Oct. 1994.