Adaptive Enhancement of Fourier Spectra

Venkatesh R. Chari and Carol Y. Espy-Wilson, *Member, IEEE*

***Abstract*— An adaptive enhancement procedure is presented which emphasizes continuant spectral features such as formant frequencies, by imposing frequency and amplitude continuity constraints on a short-time Fourier representation of the speech signal. At each point in the time-frequency field, the direction of maximum energy correlation is determined by the angle of a linear window at which the energy density within it is closest in magnitude to the point under consideration. Weighted smoothing is then performed in that direction to enhance continuant features.**

I. Introduction

An adaptive enhancement procedure which emphasizes continuant spectral features such as formant frequencies was developed during the implementation of a new formant tracking technique. Continuity in frequency and amplitude is one of the strongest constraints that can be relied upon in tracking formants. Since the articulators cannot move much in a short time interval, formant frequencies and amplitudes in one frame[1] would be expected to be close to their values in adjacent frames. The peaks corresponding to formant frequencies can thus be called "continuant" spectral features[2].

In addition to these continuant spectral features, the short-time Fourier spectrum (STFS) exhibits harmonic structure and artefacts due to windowing effects, both of which are undesirable for peak picking. The use of short duration windows reduces the prominence of the harmonics. The artefacts, however, are not subject to the continuity constraints like the formant peaks are, and are dependent on the local, short-time properties of the speech signal and the window. Thus, while the continuant features show only a slight shift from one frame to the next as the articulators slowly change position, the artefacts exhibit gross alteration in character. This can be seen in Fig. 1(a) from the sequence of short-time Fourier spectra for the vowel /e/ (as in the word "bait"), computed at intervals of 2 ms with an 8 ms Hamming window. The desired spectral profiles or envelopes, determined by the vocal tract parameters and devoid of the excitation and windowing influences, are shown in Fig. 1(b) (these were generated by the algorithm described later in this section). We now describe an algorithm for extracting the spectral envelope by emphasizing continuant features and de-emphasizing others, thereby applying continuity constraints in frequency and amplitude.

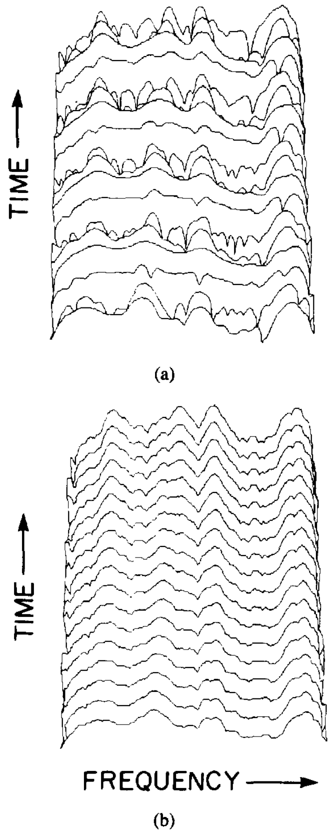


Fig. 1. (a) Short-time Fourier spectra of consecutive, overlapping segments of speech in the vowel /e/, as in the word "bait"; (b) the spectral envelopes generated by adaptively enhancing the short-time Fourier spectra.

II. Adaptive Enhancement Technique

The adaptive enhancement technique uses as its input the STFS of the speech signal. This is the squared magnitude of the decimated STFT and is computed as

$$a(n,k) = |X(nL,k)|^2 \tag{1}$$

where $X(nL,k)$ is the decimated STFT [1]. The STFS can be visualized as a 3–D representation of the signal stored in a 2-D array. The dimensions of the array are discrete time and frequency and are indexed by n and k, respectively. The value of each element is the energy at that discrete time

Manuscript received August 12, 1992; revised May 27, 1994. This work was supported by NSF Grant 9296032. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amro E. Jaroudi.

V. R. Chari was with the Department of Electrical, Computer and Systems Engineering at Boston University. He is currently with Technology for Independence, Inc., Boston, MA 02215 USA.

C. Y. Espy-Wilson is with the Department of Electrical, Computer and Systems Engineering, Boston University, Boston, MA 02215 USA.

IEEE Log Number 9406778.

[1] We define a frame to be an instant in time at which the analysis window of the short-time spectrum is centered. Typical window length would be 8 ms, with a 2 ms interval between frames.

[2] We define a spectral feature as a peak or valley in the spectrum.

1063–6676/95$04.00 © 1995 IEEE

and frequency. This 3–D representation corresponds to the spectrogram and can be plotted in two dimensions, with time and frequency axes and with energy being represented by the darkness of the point.

Before adaptive enhancement, the speech signal is segmented into voiced phonetically contiguous units to prevent spectral discontinuities from affecting the enhancement procedure. The segmentation procedure consists of three stages. The first stage separates voiced regions from unvoiced ones based on the energy in the 100–300 Hz range, and is similar to the one used in [2]. The second stage uses the peak value of the normalized cross-correlation function applied to the linear prediction residual of the speech signal [3] to determine the frames where abrupt discontinuities in formant trajectories occur within voiced regions. This procedure helps to ensure that spectral characteristics of one phonetic segment do not effect those of adjacent segments. The third stage provides pointers to regions in which the vocal tract is open and those in which it is constricted. This stage is also described in [2].

Once segmentation is complete, the adaptive enhancement algorithm first identifies continuant spectral features by finding the correlation between spectral features in adjacent frames. Continuant features like formant frequencies will have higher correlation than others. We now describe the algorithm to identify features that have the highest correlation with a point $P_{n,k}$ in the time-frequency (TF) field.

At each point $P_{n,k}$ in the TF field, we define a linear, rectangular window $w'(n)$ of length R points, centered on the point $P_{n,k}$. R is an odd integer and the window is symmetric about $P_{n,k}$. Fig. 2 depicts a 7-point window oriented at an angle $\theta = 3\pi/4$ radians with the abscissa. The average energy within the window is then computed as

$$d_{n,k}(\theta) = \frac{1}{R} \sum_{r=-\infty}^{\infty} a(n + x_{r,\theta}, k + y_{r,\theta}) w'(r) \qquad (2)$$

where $a(n,k)$ is defined in (1) and

$$x_{r,\theta} = (\text{int})\, r\cos\theta \qquad (3)$$

$$y_{r,\theta} = (\text{int})\, r\sin\theta \qquad (4)$$

and (int) represents the integer part. The time-frequency cells contained in the window of Fig. 2 are $P_{n-3,k+3}$, $P_{n-2,k+2}$, $P_{n-1,k+1}$, $P_{n,k}$, $P_{n+1,k-1}$, $P_{n+2,k-2}$, and $P_{n+3,k-3}$. Note that it is possible that the quantization in (3) and (4) may result in a point being used more than once in a given window. However, it was empirically determined that this occurred so infrequently as to be insignificant. The window is rotated in the TF plane through discrete values of θ which are restricted to be within 45° on either side of the abscissa. The average energy $d_{n,k}(\theta)$ is computed as a function of θ. The angle $\hat{\theta}_{n,k}$ at which the average energy in the window is the closest to the energy at the point $P_{n,k}$ is then given by

$$\hat{\theta}_{n,k} = \underset{\theta}{\operatorname{argmin}} \mid a(n,k) - d_{n,k}(\theta) \mid \qquad (5)$$

and is the direction of maximum energy correlation. That is, the spectral features in the direction specified by $\hat{\theta}_{n,k}$ bear the greatest similarity to the spectral feature at point $P_{n,k}$. Now that the direction of maximum energy correlation has been identified, the next step is to apply the continuity constraint by performing a weighted smoothing on the point $P_{n,k}$ with the points in the window at the angle $\hat{\theta}_{n,k}$. The new value of the energy at the point $P_{n,k}$ is then given by

$$a'(n,k) = \sum_{l=-\infty}^{\infty} a(n + x_{l,\hat{\theta}_{n,k}}, k + y_{l,\hat{\theta}_{n,k}}) w''(l) \qquad (6)$$

where $w''(n)$ is an R point sequence of binomially distributed weights, symmetric about $P_{n,k}$. The new smoothed value of energy $a'_{n,k}$ at point $P_{n,k}$ is stored in a separate data array so that it does not affect computations at other points. This enhancement procedure is repeated for all points in the TF field that have been identified as belonging to a phonetically contiguous voiced segment by the segmentation procedure.

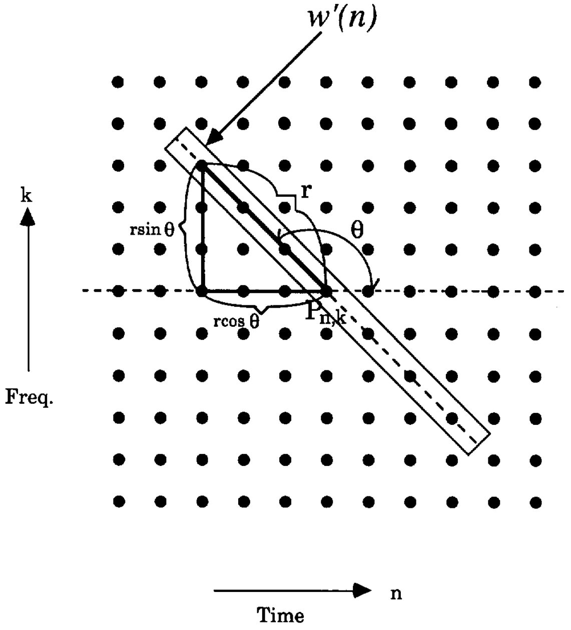


Fig. 2. Enlarged view of the time frequency field.

The adaptive enhancement procedure finds the direction of maximum correlation between spectral features in adjacent frames and then smoothes in this direction to further emphasize the correlation. Continuant spectral features are highly correlated from frame to frame and are emphasized to a greater extent than noncontinuant ones. The procedure thus serves to impose continuity constraints on the entire spectral envelope rather than just the peaks. Consequently, most of the spurious, noncontinuant features are smoothed away while genuine formant peaks remain. In this way, the adaptive enhancement technique is closer to the analysis-by-synthesis technique [4], [5] and retains its advantage of low susceptibility to spurious phenomena. This application of continuity is in contrast to existing peak-picking techniques which first pick candidates from the short-time envelope and then apply continuity constraints on them to assign them to formant slots. This method is an indirect process involving

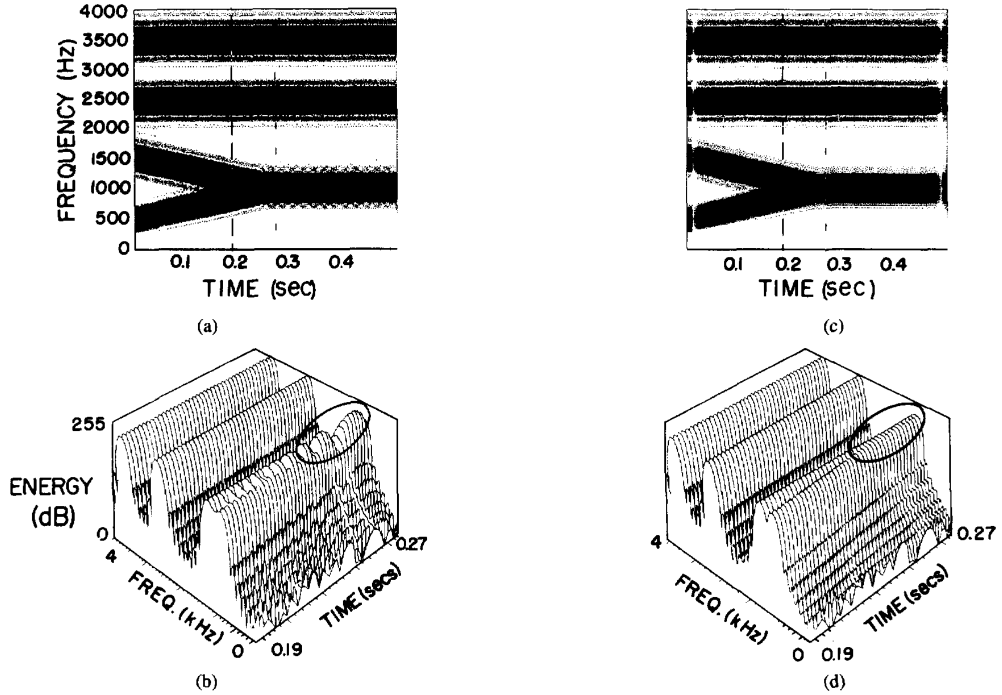


Fig. 3. (a) Spectrogram of 0.5 s of synthetic speech; (b) 3-D plot of the section from 0.186 to 0.268 s; (c) spectrogram of the segment after adaptive enhancement; (d) 3-D plot of the section from 0.186 to 0.268 s after adaptive enhancement.

two decision-making stages. An error in the first peak-picking stage, induced for instance by picking a spurious peak, is sometimes irrecoverable.

The ability of the adaptive enhancement technique to de-emphasize less continuant peaks is important in dealing with nasalized vowels where an extra peak is present. Generally, the nasal formant is shorter in its duration than oral formants, so that it does not persist over the entire phonetic segment. Consequently, the nasal formant may be emphasized less than its oral counterpart and may therefore be smoothed away. There are, however, instances when the nasal formant is of appreciable duration. In such cases, the peak train corresponding to the nasal formant will have to be eliminated in the next stage by calling upon higher knowledge sources.

The adaptive enhancement helps in the resolution of merged formants, especially if the duration of the merger is less than the time during which the peaks were distinct. This ability to resolve close formants is again due to the application of continuity constraints on the entire spectral profile. When operating on an area of the TF field where the formants are close together and appear merged, the correlating procedure identifies different points in the vicinity of the peak as being correlated to the distinct peaks in the adjacent time frames. The enhancement stage then smooths these points with the distinct peaks and emphasizes them over others. This procedure is illustrated by considering the following test case.

A 0.5-s segment of synthetic speech was generated by the addition of four sinusoids at frequencies of 500, 1500, 2500, and 3500 Hz to represent the first four formant frequencies. While the higher two formants remained constant in frequency for the entire duration of the segment, the lower two were swept with time. The sinusoid at 500 Hz was linearly swept up at a rate of 2 kHz/s until it reached 1000 Hz at 0.25 s. It then remained constant in frequency till the end of the segment. The sinusoid at 1500 Hz was linearly swept down in frequency at 1800 Hz/s until it reached 1050 Hz at 0.25 s. It then remained at a constant frequency of 1050 Hz for the rest of the segment. Fig. 3(a) shows a spectrogram of the composite signal. Since the sampling frequency for the signal was 8 kHz, an 8 ms analysis window for the STFT contains 64 samples. The STFT has a maximum frequency resolution of 62.5 Hz and, therefore, cannot resolve the sinusoids at 1000 and 1050 Hz beyond 0.25 s. This lack of distinction between the two peaks can be seen from the 3-D plot of the portion of the segment from 0.186 to 0.268 s in Fig. 3(b). Most frames beyond 0.25 s appear to have a single peak below 2 kHz.

The adaptive enhancement algorithm was then applied to the synthetic speech segment. Fig. 3(c) shows the spectrographic representation after the enhancement. While not much change is discernible from the spectrogram, the 3-D plot of the section from 0.186 0.268 s in Fig. 3(d) clearly shows the effect of the adaptive enhancement algorithm. It can be readily seen that the sinusoids at 1000 and 1050 Hz that appeared as one peak in the STFS are resolved into two distinct peaks for almost the entire segment. Thus, the enhancement technique enables formants that are merged in the original STFS to be resolved. In cases where the formants are merged through their entire length, the enhancement technique is not as effective since it

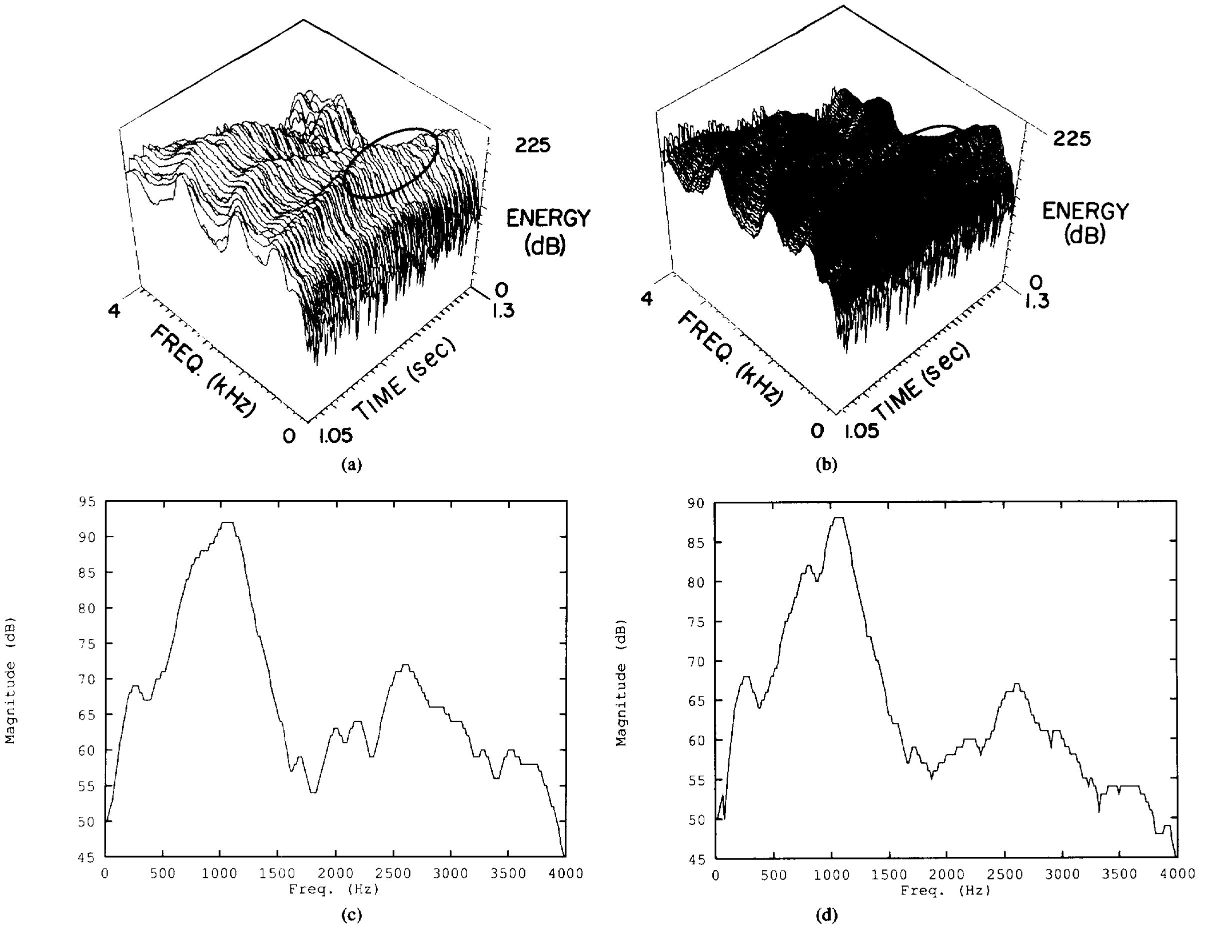


Fig. 4. (a) 3–D plot of the spectrogram of a segment of natural speech; (b) 3-D plot after adaptive enhancement; (c) spectrum taken at 1.25 s from the original spectrogram shown in (a); (d) spectrum taken at 1.2 ms from the adaptively-enhanced spectrogram shown in (b).

is unable to identify any distinct peaks with which to correlate the merged peak.

A similar test was conducted on a segment of natural speech in which F1 and F2 were merged. The 3–D plots of the spectrograms of the segment before and after adaptive enhancement are shown in Fig. 4(a-b), respectively. While F1 and F2 appear as a single peak towards the end of the plot of the original spectrogram, they appear as two distinct peaks in the plot of the spectrogram after adaptive enhancement. The difference before and after adaptive enhancement can also be seen from a comparison of the time slices at 1.25 s. The original spectrum in Fig. 4(c) consists of a prominent F2 peak at about 1000 Hz, but F1 shows up as a shoulder resonance. In contrast, the adaptively enhanced spectrum in Fig. 4(d) consists of two distinct peaks for both F1 and F2.

III. Conclusion

In this paper we have presented an adaptive enhancement algorithm which operates on a short-time Fourier representation of the speech signal to avoid problems associated with model-based spectra. The continuity of formants in frequency and amplitude was exploited to detect continuant spectral features that exhibited spectral correlation over extended durations. The enhancement process then emphasized such features over transient spectral features to yield a spectral envelope that was more conducive to peak-picking. Experiments on synthetic and natural speech demonstrated the efficacy of the adaptive enhancement technique in sharpening formant peaks and separating formants that are merged for a portion of their length.

References

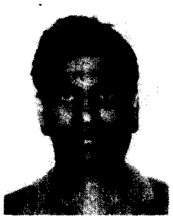
[1] S. H. Nawab and T. F. Quatieri, "Short-Time Fourier Transform," *Advanced Topics in Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[2] C. Y. Espy-Wilson, "A feature-based approach to speech recognition," *J. Acoust. Soc. Am.*, vol. 96. pp. 65–72, 1994.

[3] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," *Proc. 1993 IEEE ICASSP* (Boston), April 1983, pp. 1352–1355.

[4] C. G. Bell *et al.*,"Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Am.*, vol. 33, pp. 1725–1736, 1961.

[5] J. P. Olive, "Automatic formant tracking by a Newton-Raphson technique," *J. Acoust. Soc. Am.*, vol. 50, pp. 661–670, 1971.



Venkatesh R. Chari was born in New Delhi, India, in 1967. He received the Bachelor of Engineering degree from Maharaja Sayajirao University in 1990 and the M.S. degree from Boston University in 1992, both in electrical engineering.

He is currently with Technology for Independence, Inc., Boston, MA, where he is involved in the development of assistive technology for the blind and visually impaired. His research interests include speech synthesis, coding, and recognition techniques and their application in the design of embedded systems for use in assistive devices.

Carol Y. Espy-Wilson (S'81–M'90) was born in Atlanta, GA in 1957. She received the B.S. degree from Stanford University, Stanford, CA, in 1979, and the M.S., E.E., and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1981, 1984, and 1987, respectively, all in electrical engineering.

From 1987 to 1988, she was a Postdoctoral Fellow at the Research Laboratory of Electronics (RLE), MIT, and was a part-time member of technical staff in the Linguistics Research Department at AT&T Bell Laboratories, Murray Hill, NJ. From 1988 to 1990 she was a Research Scientist in RLE, MIT. Currently, she is an Assistant Professor in the Electrical, Computer and Systems Engineering Department at Boston University. Her research interests include speech communications with a focus on speech recognition and digital signal processing.

Dr. Espy-Wilson is a recipient of the Clare Boothe Luce Professorship, and is a member of ASA.