CHARACTERIZING FORMANTS THROUGH STRAIGHT-LINE APPROXIMATIONS WITHOUT EXPLICIT FORMANT TRACKING

S. Seneff

Rm 36-549, Research Laboratory of Electronics, M.I.T.,
Cambridge, MA. USA 02139.

A new method for representing the formants of sonorant speech sounds is described. The method collapses the two-stage process of (1) formant tracking and (2) abstraction of rates and directions of formant movements into a one-step process of directly assigning straight-line segments to the resonance contours in the frequency-time space. The method resembles techniques used in vision research [1], and is also motivated by observations of specialized frequency-modulation detectors in the central auditory system [4]. The computational procedures are straightforward, leading to a description of the formant information for a given vowel by a list of oriented straight-line segments. The line segments are not assigned to particular formants, such as F_2. Instead, the recognition process is hypothesis-driven. For each vowel or diphthong to be recognized, a short description of expected ranges of frequency and orientation in the time-frequency dimensions for the first two formants is given. Feasibility of the method is demonstrated by applying it to the specific task of recognizing the vowels and diphthongs of American English in restricted context, spoken by multiple speakers.

OVERVIEW

It is generally accepted that the frequencies of the formants, particularly the first two formants, are the most important information leading to the identification of vowels. Formant movements are also necessary for identifying diphthongs and semivowels. As a result, a number of investigators have attempted to develop formant tracking algorithms, which assign spectral peaks to specific formants, such as F_1, F_2 and F_3. Once the formant tracks are available over time, it is possible to develop algorithms that detect high-level features, such as a rising formant over the second half of a vowel.

Our approach is to represent the formant information directly by a collection of straight line segments, thus bypassing the stage of formant tracking. The formant patterns are described by oriented lines which often overlap in time and/or frequency, and which collectively provide sufficient information for identification of the phonetic content. These line segments lead naturally to descriptions such as "rising formant", with the slope of the line conveying the degree of rise.

The spectral representation, the "pseudo spectrum," from which the line segments are abstracted is obtained using an auditory-based signal processing method, as described in [2]. The method typically yields enhanced peaks at formant frequencies with smooth transitions over time. For voices with a high fundamental frequency, the individual harmonics of the pitch are often resolved below the first formant, thus making it very difficult to track F_1 in the traditional way.

LINE FORMANT EXTRACTION PROCESS

The process to obtain a list of straight-line segments describing the formant patterns in a given sonorant segment of speech is illustrated in Figure 1. The pseudo spectrogram for the word "Burt", spoken by a male speaker, is shown in Part (a) of the Figure, with the frequency axis represented on a Bark scale. A nonlinear filter-and-quantize procedure defines "On" and "Off" contour regions in time and frequency, shown in Part (b). Each robust peak in a given pseudo spectral cross-section is allowed to vote for a best-fit line segment passing through its time-frequency location, restricted to stay within an "On" region, and oriented in one of 11 specified directions. The votes of the robust peaks are accumulated in a list giving information about the orientation, center-points in time and frequency, duration, and mean amplitude of each line.

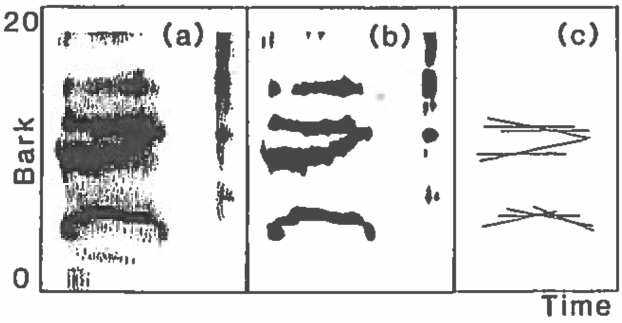

Figure 1: Illustration of Line-formant Abstraction Process (a) Pseudo spectrogram for word "Burt"; (b) One-bit enhanced spectrogram defining allowable regions for line segments; (c) Resulting line segments describing formants of vowel.

The next step is to consider collectively the list of candidate lines over a time interval defined by the unknown vowel's extent. Usually, several peaks vote for the same line or very similar lines. A heuristic algorithm was developed to collapse the list of lines into a new list, with "equivalent" lines merged into a single representative, which includes a count of the number of votes being merged. Finally, the list is further pruned, and line segments that appear to be insignificant are discarded. Elimination is based on threshold requirements for the number of votes, the minimum allowable duration, and the mean amplitude. The line segments that remain after pruning in the example are shown in Part (c) of the Figure.

The final step is to convert the list of line segments into a fuzzy descriptor format. The temporal extent of a given line is converted to a verbal description of its extent relative to the vowel end points, such as "first half". Similarly, the strength and orientation of the line are quantized to a small set of possibilities. Only the center frequency is retained as a number. Table 1 lists allowable categories for each item.

Orientation		Temporal		Strength
Rapid Rise	Rapid Fall	At Start	At End	Strong
Rising	Falling	First Half	Second Half	Medium
Slight Rise	Slight Fall	In Middle	Throughout	Weak
Steady				

Table 1: Categories for descriptors of line formants.

VOWEL RECOGNITION STRATEGY

The line formant representation was applied in a speaker-independent recognition task for the following 16 vowels and diphthongs of English, restricted to /bVt/ context: /i, e, yu, I, ɛ, æ, ɑ, ɔ, o, ʌ, ʊ, u, ɑ^{y}, ɑ^{u}, ɔ^{y}, ɝ/. The only step used for speaker normalization was to reference the center frequency *in*