STOP CLASSIFICATION USING DESA-1 HIGH-RESOLUTION FORMANT TRACKING

J. T. Foote *D. J. Mashao* *H. F. Silverman*

Laboratory for Engineering Man-Machine Systems
Division of Engineering
Brown University, Providence, RI 02912

ABSTRACT

Recent work has verified that the second-formant frequency (F2) and its change in the vowel immediately preceding a stop consonant are usually sufficient to discriminate between labial, palatal, and alveolar stops, even in the absence of the stop burst information. Informal listening tests using truncated samples indicate that humans can discriminate the three stops on the basis of the preceding vowel alone. Typical quasi-stationary analyses like LPC and DFT filterbanks may not have sufficient time-frequency resolution to detect the rapid F2 variations, and therefore a valuable source of stop classification information is being overlooked. This paper shows the results of using the DESA-1 quadratic frequency estimator to determine the frequency and rate of change of F2. It is shown for different vowel environments that the DESA-1 algorithm can extract sufficient information to classify stops from vocalic data. The performance is demonstrated to be superior to a formant tracker using a more conventional pitch-synchronous LPC analysis.

1. THE DESA-1 ALGORITHM

The Discrete Energy Separation Algorithm (DESA-1) recently presented by Maragos, Kaiser, and Quatieri [1, 2] is based on the work of H. Teager. The DESA-1 algorithm provides a simple and elegant method of estimating the amplitude and frequency of a sinusoid subject to amplitude or frequency modulation. Using the notation of [2], the Teager energy-tracking operator is defined as

$$\Psi_c[x(t)] \triangleq \dot{x}^2(t) - x(t)\ddot{x}(t) \tag{1}$$

where $\dot{x}(t)$ and $\ddot{x}(t)$ are the first and second time derivatives of $x(t)$. Given a sinusoid with amplitude modulation $a(t)$ and instantaneous frequency $\omega(t)$,

$$x(t) = a(t)\cos(\int_0^t \omega(\tau)d\tau) \tag{2}$$

the Teager energy operator estimates the squared product of the instantaneous amplitude and frequency, under certain reasonable conditions (essentially, modulation frequencies must be slow with respect to the carrier).

$$\Psi_c[x(t)] \approx a^2(t)\omega^2(t) \tag{3}$$

Applying the operator to the *derivative* of the AM-FM signal yields

$$\Psi_c[\dot{x}(t)] \approx a^2(t)\omega^4(t). \tag{4}$$

Clearly, the instantaneous amplitude and frequency may be found from the Teager energy estimates of the signal and its derivative by solving Equations (3) and (4) for a and ω.

In the discrete case, time derivatives may be approximated by time differences. The discrete-time counterpart of the Teager operator $\Psi_c[x(t)]$ becomes:

$$\Psi_d[x(n)] \triangleq x^2(n) - x(n-1)x(n+1) \tag{5}$$

A discrete-time AM-FM sinusoidal signal having amplitude modulation $a(n)$, carrier frequency Ω_c, modulation frequency Ω_m, and modulation function $q(n)$, can be expressed as (ignoring an arbitrary phase constant)

$$x(n) = a(n)\cos(\phi(n)) = a(n)\cos(\Omega_c n + \Omega_m \int_0^n q(k)dk) \tag{6}$$

and has instantaneous frequency $\Omega_i(n)$

$$\Omega_i(n) = \frac{d\phi(n)}{dn} = \Omega_c + \Omega_m q(n) \tag{7}$$

It has been shown [3] that the discrete-time Teager energy operator estimates the following function of the instantaneous amplitude and frequency:

$$\Psi_d[x(n)] \approx a^2(n)\sin^2(\Omega_i(n)) \tag{8}$$

again under the reasonable assumptions that the modulation functions are slowly varying with respect to the carrier frequency Ω_c.

Because the forward and backward differences are not symmetric, a symmetric approximation to the operative of the derivative of (6) may be found by averaging the operative results of the backward difference $y(n) = x(n) - x(n-1)$ and the forward difference $z(n) = x(n+1) - x(n)$. In this case,

$$\frac{1}{2}(\Psi[y(n)] + \Psi[z(n)]) \approx 4a^2(n)\sin^2[\frac{\Omega_i(n)}{2}]\sin^2[\Omega_i(n)] \tag{9}$$

(again as shown in [3]). The DESA-1 algorithm yields an estimate of the instantaneous amplitude and frequency by solving Eqs. (8) and (9) for $a(n)$ and $\Omega_i(n)$ as follows:

$$G(n) \triangleq 1 - \frac{\Psi[y(n)] + \Psi[z(n)]}{4\Psi[x(n)]} \tag{10}$$

$$\Omega_i(n) = \cos^{-1}[G(n)] \tag{11}$$

$$|a(n)| = \sqrt{\frac{\Psi[x(n)]}{1 - G^2(n)}} \tag{12}$$

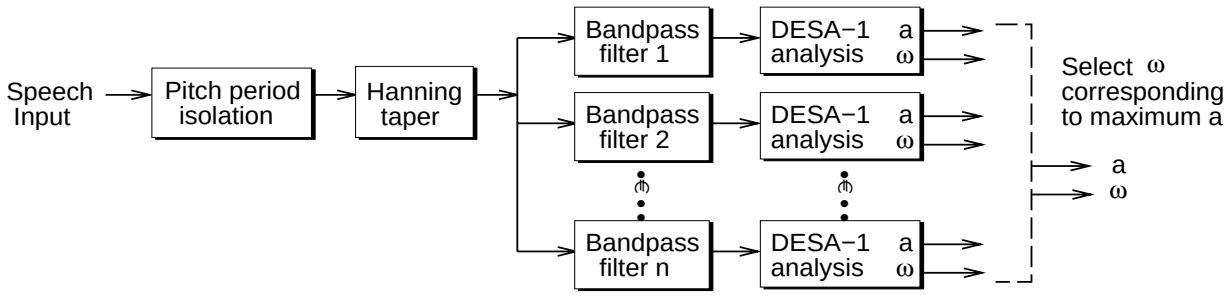

Figure 1: Frequency estimation block diagram

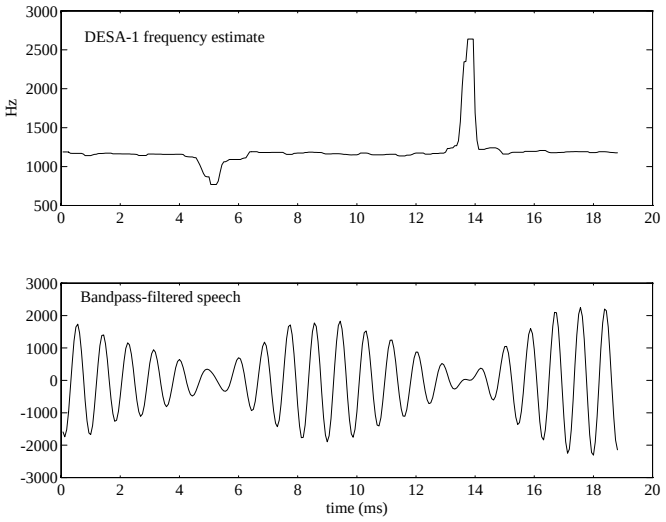

Figure 2: Frequency artifacts at pitch period boundaries

2. USING DESA-1 FOR FORMANT TRACKING

Real speech data, of course, is more complex than the simple AM-FM sinusoids of the previous section. However, the AM-FM signal is still a reasonable model for individual formants, which may be isolated by appropriately filtering the speech signal. Unfortunately, straightforward application of the DESA-1 algorithm is problematic because interactions between the filter and neighboring pitch periods causes undesirable artifacts in the frequency estimates.

Towards the end of a pitch period, when the glottal energy is mostly dissipated, the filter starts to overlap the start of the *next* glottal impulse. The filtered signal is in the right frequency range, but is usually not in phase with the current pitch period. Figure 2 illustrates this phenomenon. The bandpass-filtered speech signal has pitch period boundaries near 5 and 14 milliseconds. The frequency is roughly constant across the pitch periods, but there are phase discontinuities at the boundaries, which give rise to the frequency estimate spikes.

2.1. Pitch Period Detection

At first glance, it would seem that merely ignoring the pitch boundary regions would be sufficient to reduce the frequency artifacts, but variable pitch energy, especially when approaching a stop, means that artifacts can occur anywhere in the pitch period, not just in the low-energy pitch-boundary regions of Figure 2. This problem may be overcome by processing pitch periods in isolation.

Fortunately, the formant-tracking algorithm does not require the precise glottal-closure instant detection of [4]. Rough pitch-period boundaries were detected by bandpass filtering the speech signal in the first formant frequency range. Regions with a smoothed Teager energy of greater than 130% of the mean energy were regarded as valid pitch intervals for further processing. The Teager energy was used rather than the squared-signal energy because it is more likely to be small at the problematic pitch period boundaries. This is because the Teager energy, being proportional to the square of the frequency as well as the amplitude, is larger at the beginning of the pitch period when there are more high-frequency components from the glottal function.

2.2. Pitch-Synchronous Formant Tracking

Once isolated, the pitch interval was tapered with a Hanning (raised cosine) window of the same length to reduce truncation effects. The tapered data were zero-padded and fed into a filterbank of four bandpass filters. Depending on the vowel environment, the filters spanned the frequency range 1100–1600 Hz (the first experiment below) or the range 1400-2000 Hz for the vowel /æ/. Each filterbank section consisted of a 257-point optimal FIR filter having a 200 Hz passband width; adjacent filters overlapped by 100 Hz.

DESA-1 analysis was then performed on each filter output, which resulted in a time-varying frequency and amplitude estimate for the duration of the pitch period. To find the general region of the second formant, the amplitude estimates were summed for each frequency range. The DESA-1 frequency estimate from the filter with maximum summed amplitude was chosen as a good estimate of F2 over the pitch period. The relatively long filter serves to smooth the frequency estimate, so no additional smoothing (like the median filter of [2]) is required. Note that it is not necessary to compute the DESA-1 analysis for every filterbank output: in time-critical applications, a simpler energy estimate could certainly be used. In studies of consonant-vowel transitions [5], it was found that formant trajectories are well-approximated by straight lines. The rate of change of F2 is thus determined by the slope of the LMS best linear fit to the F2 estimates over the last few pitch periods approaching a stop.

3. EXPERIMENTS

It has long been known that the formant transitions between vowels and stops are important cues to the place of stop articulation. [6, 7, 8, 9]. Informal listening tests [10] indicate that humans can discriminate between the unvoiced stop consonants /p/, /t/ and /k/ on the basis of the preceding vowel alone. Given recorded speech data of "pupper",

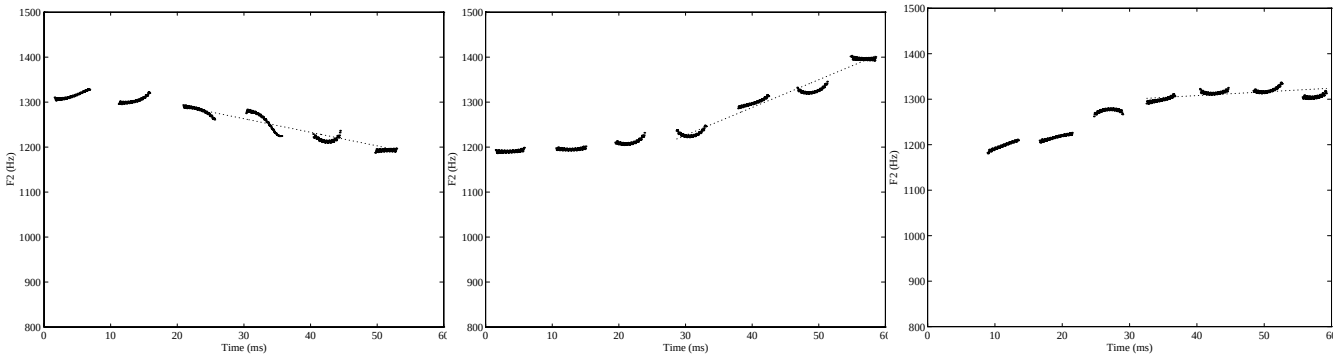

Figure 3: Estimated F2 and best linear-fit slope in vowel before P (left), T (center), and K (right).

"putter," and "pucker," untrained listeners were played the initial "puh" and asked to judge whether the (unheard) stop was a /p/, /t/ or /k/. The subjects chose the correct stop with better than 80% accuracy, indicating that coarticulation effects in the preceding vowel give valuable information about the stop.

Three instances of the pseudo-words "upper," "utter," and "ucker" spoken by 20 male talkers were sampled at 16 kHz. Talkers were instructed not to flap the /t/, i. e. "udder" was not permitted for "utter." The vowels before the unvoiced stop consonants /p/, /t/ and /k/ were automatically extracted and served as the corpus for the first experiment. When examined over the last four pitch periods before the stop, the frequency and slope of F2 are good indicators of the stop type. Figure 3 shows the results of the F2 estimation procedure on the vowel /ə/ before the stops /p/, /t/ and /k/. Labial stops (/p/) are characterized by a relatively low mean F2 frequency with a flat or negative formant slope, alveolar stops (/t/) have a high F2 frequency and a large positive slope, while palatal stops (/k/) have an intermediate frequency with slightly positive slope. Figure 4 shows the mean F2 value plotted against the F2 slope for the last four pitch periods. The different stops are reasonably well-clustered. Note that the distinctive slope occurs in less than 40 milliseconds, which is about the window length used in typical speech recognition systems. Clearly, long analysis windows may obscure important time-varying information.

To compare with conventional techniques, the same isolated-pitch data were used in an LPC-based formant tracker. The speech data were first lowpass filtered to 3200 Hz, using a zero-phase FIR filter to preserve time alignment. An 8-pole, autocorrelation LPC analysis was performed on the tapered pitch intervals determined as in the DESA-1 experiment. The second formant was considered to be the pole frequency of the largest-magnitude pole pair found between 1100 and 1600 Hz. If no poles were found, pairs between 1000 and 2000 Hz were considered. Figure 5 shows the mean F2 of the last four pitch periods plotted against the slope of the best LMS linear fit to the pitch-synchronous LPC F2 estimate, while Figure 4 shows the feature values determined by the DESA-1 method. The difference is clear: the DESA-1 estimates are reasonably clustered and classification could be performed, while the LPC method does not discriminate well between the classes and suffers from particularly poor slope estimates (note the different Y–axis scales between the two figures).

4. DISCUSSION

Interpretation of these results is aided by numerous studies of consonant-vowel formant transitions in the literature. It is generally agreed that lip-rounding causes F2 to drop, because it is acoustically equivalent to increasing the length of the resonant cavity [11]. This agrees with the the slightly negative F2 slope seen for labial stops, and results reported elsewhere [7, 5]. The situation for alveolar and palatal stops is not as straightforward. From perceptual studies using synthetic speech [6], it has been hypothesized that alveolar stops have a "locus" near 1800 Hz. That is, F2 will move from the steady-state vowel location to somewhere near 1800 Hz in the stop. This is consistent with the distinctly positive slopes found here. However, there is some controversy whether invariant "loci" really exist in natural speech, especially as most of the studies involved only a single talker and did not investigate inter-talker variations.

Preliminary studies in other vowel environments indicate that similar cues are present, though they may not be as distinct. Figure 6 shows the that reasonable discrimination was obtained for one talker in the vowel environment "apesh," "atesh," and "ackesh," (initial vowel (/æ/ as in "had"). A similar experiment across 20 talkers revealed much poorer separation than the previous "upper"-"utter"-"ucker" environment, especially in the slope dimension. It has been hypothesized that F2 of /æ/ is quite near the reported "locus" of alveolar stops, thus slopes are smaller in magnitude and may be variable due to natural inter-speaker variations. For instance, if a talker has a naturally high F2, the slope for an alveolar stop may be negative, and thus resemble a labial stop. Work is underway in extracting cues that are invariant across vowel environments.

5. SUMMARY

The DESA-1 algorithm is a computationally simple and robust way of detecting rapid formant variations. The results presented here demonstrate the advantage of nonlinear, nonstationary analysis over conventional techniques: The DESA-1 method yields results comparable to the time-varying LPC method of [4], and can effectively extract stop classification information for speech recognition applications. In addition, the DESA-1 method avoids the iterative estimation procedures of [4], and is computationally simple enough to be performed in real time.

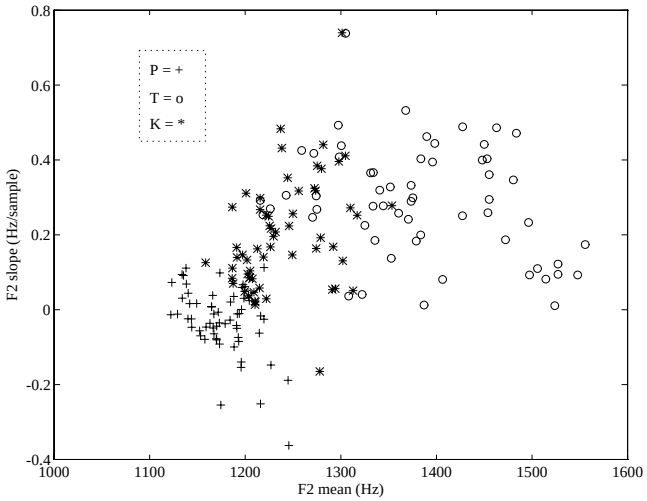

Fig. 4: F2 slope vs. F2, 20 talkers, DESA-1 results.

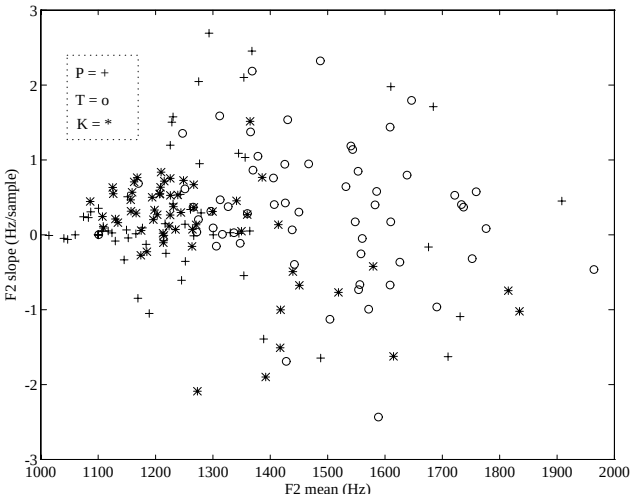

Fig. 5: F2 slope vs. F2, 20 talkers, LPC results.

6. ACKNOWLEDGEMENTS

The authors are indebted to James Kaiser for his helpful comments, and to Krishna Nathan for his work on the formant-slope phenomenon. We also thank N. Rex Dixon for his advice. This work was funded in part by NSF grant #MIP–9120843.

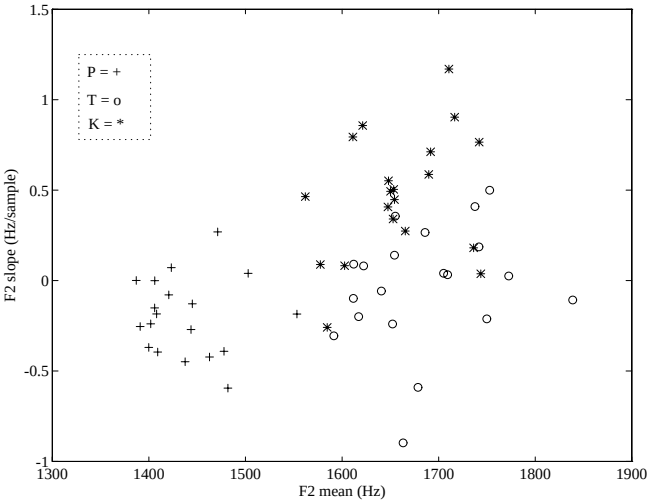

Fig. 6: DESA-1 F2 slope vs. F2, single talker, vowel /æ/.

REFERENCES

[1] P. Maragos, J. Kaiser, and T. Quatieri. Speech nonlinearities, modulations, and energy operators. In *Proceedings 1991 ICASSP*, pages 421–424. IEEE, May 1991.

[2] P. Maragos, J. Kaiser, and T. Quatieri. On separating amplitude and frequency modulations using energy operators. In *Proceedings 1992 ICASSP*, volume II, pages 1–4. IEEE, March 1992.

[3] P. Maragos, J. Kaiser, and T. Quatieri. On detecting amplitude and frequency modulations using energy operators. Technical Report 91-6, Harvard Robotics Lab., 1991.

[4] K. Nathan and H. Silverman. A time-varying analysis method for rapid transitions in speech. *IEEE Trans. SP*, 39(4):815–824, April 1991.

[5] D. Kewley-Port. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *J. Acoust. Soc. Am.*, 72(2):379–389, August 1982.

[6] P. Delattre, A. Liberman, and F. Cooper. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.*, 27(4), March 1955.

[7] G. Fant. *Speech Sounds and Features.* MIT Press, Cambridge, Mass., 1973.

[8] Piero Demichelis and Renato DeMori. Computer recognition of plosive sound using contextual information. *IEEE Trans. ASSP*, 31(5):369–377, April 1983.

[9] M. Bush, G. Kopec, and V. Zue. Selecting acoustic features for stop consonant identification. In *Proceedings 1983 ICASSP*, volume 2, pages 742–745. IEEE, 1983.

[10] D. Mashao. Perception of stop consonants from the preceding vowel. Technical Report 109, LEMS, Division of Engineering, Brown University, Providence RI, 1992.

[11] Philip Lieberman and Sheila Blumstein. *Speech Physiology, Speech Perception, and Acoustic Phonetics.* Cambridge University Press, Cambridge, UK, 1988.