DYNAMIC VS. STATIC SPECTRAL DETAIL IN THE PERCEPTION OF GATED STOPS

Michael Kiefte and Terrance M. Nearey

Dept. Linguistics, University of Alberta

ABSTRACT

In order to assess the importance of dynamic spectral information within the first few milliseconds following oral release for the identification of prevocalic stop consonants, 23.75 ms gated CV syllables were presented to listeners for identification. In addition to these, subjects were presented with the same tokens reconstructed from their minumum phase decomposition such that they have the same long-term power spectrum as their original counterparts, but with differing internal dynamic spectral detail. Subjects' results from this experiment were then modelled with logistic regression analysis using mel cepstral coefficients with and without dynamic spectral information encoded in order to demonstrate the effect that reduced temporal information has in the context of automatic classification. Preliminary results from this experiment show that some dynamic spectral detail is used by listeners even for very short stimuli. We conclude that models of speech perception must take spectral variation over very short time frames into account.

1. INTRODUCTION

It has been suggested by Stevens and Blumstein that the configuration of the articulators within the oral tract at the moment of release for stop consonants produces an acoustic waveform that is sufficient for the categorisation of this class of speech sounds. This hypothesis has lead to the conclusion that the relevant information is contained in the first 20 ms following the release burst — possibly including the subsequent vocalic segment[1,2]. This position has been challenged numerous times on the grounds that spectrally dynamic information is additionally necessary for correct classification[e.g. 3].

However, spectrally dynamic models of stop consonant perception are based on much longer signal durations following oral release: either by a fixed length (e.g. 40 ms[3], 60 ms[4]), or variable length based on the duration of the burst[5,6]. Despite this, it is known that listeners are quite capable of identifying gated stops of 10-20 ms in duration[7] and that classific ation based on the burst alone is also statistically better than chance[8]. Therefore, the original views of Stevens and Blumstein are not entirely in conflict with later findings; it is entirely possible that the perception of short gated prevocalic stop consonants is determined by a static spectral representation of the acoustic stimulus, while longer signals introduce perceptually relevant dynamic information.

In order to test this hypothesis, a perceptual experiment using short gated CV syllables was performed in which the stimuli were manipulated in such a way that their long-term power spectrum remained unchanged from the original sampled tokens, while their internal temporally varying structure was distorted by altering the phase of their Fourier transform. Under the assumption that no spectrally dynamic information within these short stimuli are used by listeners in categorisation, we expect that this transformation will have little effect on overall performance. Conversely, a decrease in performance would indicate that some dynamic information is being exploited by listeners to ditermine the phonetic identity of these short signals.

Even if some dynamic information exists in the original stimuli and is detectible by listeners, this fact alone does not necessarily imply that it is a usable phonetic cue. Note that we are explicitly avoiding the complexities associated with auditory modelling by focusing attention on simple signal manipulations with the intention of demonstrating the phonetic properties of temporally distorted speech signals.

1.1 Minimum Phase Decomposition

In order to preserve the magnitude spectrum of the stimuli while distorting the internal temporal detail, several operations were considered:

- Linear phase reconstruction by zeroing the phase of the FFT followed by a circular shift to produce a symmetric signal
- Allpass filtering
- Calculation of the minimum phase component of a minimum phase/allpass decomposition.

The first option produced signals that were largely impulse-like and did not even resemble possible speech sounds. The second option proved to be impractical since the complexity of the allpass filter that was desired increased the length of the resultant signal significantly, making the processed signals easy to distinguish from the original natural tokens. It was then decided to construct all-pass filters on a case-by-case basis using the zeros of the z-transform of the original signals as poles for the allpass filters. This has the effect of inverting the targeted zeros. Although time consuming, this procedure proved to be remarkably effective at producing speech like sounds. It was also surprisingly stable even given the caveats for root-solving for even modestly long signals, since the poles of the allpass filter only had to approximate the zeros of the original signal in order to preserve the duration of the phase-altered stimuli. However, because zeros both inside and outside the unit circle in the z-plane were inverted indiscriminately, the resultant signal was roughly gaussian in shape and many of them sounded more like fricatives than stops.

It was therefore decided to only invert the zeros outside the unit circle resulting in a minimum phase signal. This operation may also be performed by calculating the minimum phase/allpass decomposition[9]

$$x[n] = x_{\min}[n] * x_{ap}[n]$$

where $x_{\min}[n]$ has the same magnitude spectrum as $x[n]$ but with all zeros within the unit circle while $x_{ap}[n]$ is a signal with absolute magnitude equal to 1. This resulted in signals with sharper onsets and rapidly falling waveform envelopes that are more typical of stop consonants in general. Reconstruction of the minimum phase decomposition was accomplished using a cascade direct form II implementation of allpass filters on the time-reversed signal with poles and zeros determined from the z-transform of the original signal.

The upper plot in Figure 1 shows the original 23.75 ms gated onset of the syllable /dok/ as spoken by speaker "e", while the lower plot shows the waveform of the minimum phase reconstruction. Figure 2 shows the aplitude spectrum of the original token in the upper graph and the dB difference in amplitude spectra between $x[n]$ and $x_{\min}[n]$ which is equivalent to the magnitude response of $x_{ap}[n]$.

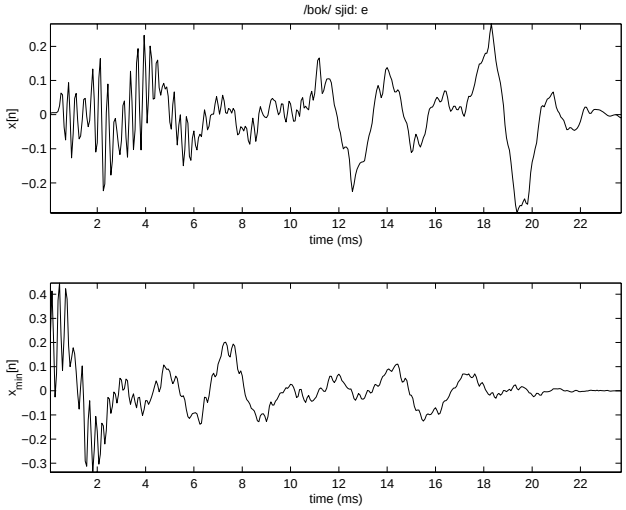

Figure 1. Waveforms for the original unaltered gated onset of /dok/ and its minimum phase reconstruction.

Most of the distortion is confined to lower amplitude high-frequency regions which were more susceptible to quantisation noise. Prior to 16 bit quantisation of the minimum phase reconstructions, the maximum absolute frequency distortion never exceeded 10^{-7} dB. Figure 3 gives the spectrographic representation of the two signals. The original token can be heard here [SOUND 898_01.WAV], and the minimum phase reconstruction here [SOUND 898_02.WAV]. In general, minimum phase reconstructions force the energy within individual frequency bands towards the onset of the signal. In this particular example, very little in the spectrographs have changed with the exception of a region of energy below 1000 Hz which is shifted towards the beginning, as well as a greater concentration of energy before 2 ms in the higher frequency regions.

3. EXPERIMENT

3.1. Stimuli

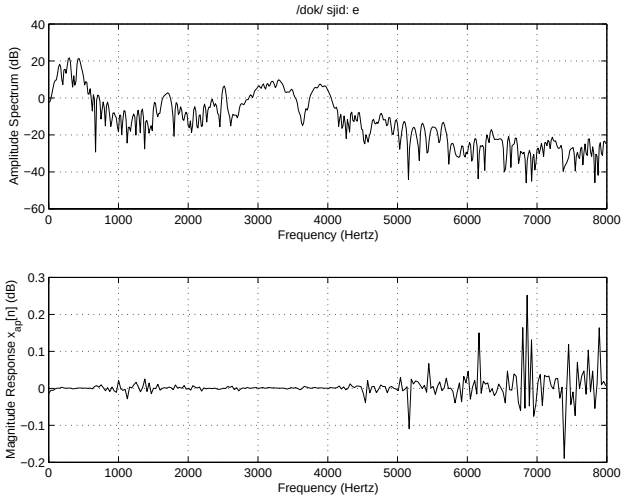

Figure 2. Magnitude spectrum of the original unaltered gated onset of /dok/ and the magnitude difference between that and its minimum phase which is equivalent to the magnitude response of the allpass filter decomposition.

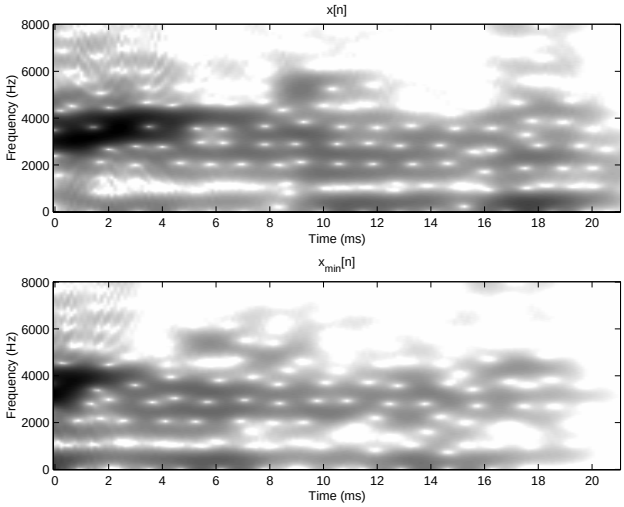

Figure 3. Spectraphic representation of original unaltered gated onset and minimum phase reconstruction of /dok/.

Natural /CVk/ tokens were recorded from 12 native speakers of Western Canadian English where C represents one of the six oral stops /p,t,k,b,d,g/ and V represents one of 10 monophthongs. The tokens were sampled at 16 kHz with 10 bit quantisation. For the purposes of this experiment, only the syllables with the vowels /e,æ,ɒ,o/ were considered, since they represented the extremes of the F1 x F2 vowel space in this dialect[10].

The signals were truncated at 23.75 ms following the onset of release burst energy and the final 5 ms was weighted with a half-hamming window. This constituted the unaltered stimulus set. The minimum phase set was generated as described above. The stimuli were played back in an anechoic room via a digital to analog converter installed in a PC. At each presentation, stimuli were played twice with an intervening 500 ms delay. An attenuator was present within the room and subjects were told that they could

adjust the control for comfort. Subjects were also able to have the stimuli repeated an unlimited number of times.

3.2. Subjects and Responses

Six volunteers were asked to serve as subjects for the experiment. Listeners were presented with a dialog box containing a button for each of the six consonants. Subjects were asked to identify each of the tokens once: both unaltered and minimum phase from each of 12 speakers, six consonants and four vowels = 576 categorisations per subject.

4. RESULTS

Overall correct responses to the onset categories was 51.1% for the unatlered stimuli as compaired with 44.4% for the minimum phase tokens. Both values are much larger than chance (16.7%). For correct place of articulation, these scores were 74.3% and 69.3% respectively as compaired with 33.3% for chance. For correct voicing distinction, 67.3% of the unaltered stimuli were classified correctly, compaired to 61.7% for the minimum phase tokens. Since the perception of the voicing distinction was much closer to chance (50%), it was decided to focus on the affect that minimum phase reconstruction had on place perception.

A generalised linear model was fit for each subjects' data for correct classification assuming an underlyingly binomial distribution[11,12]. The largest model considered included coefficients for consonant C, vowel V, a consonant by vowel diphone interaction CV, phase (whether unaltered or minimum) P and all interactions between phase and the preceding terms — PC, PV and PCV.

By minimising the Akaike Information Critereon (AIC),

$$Q = D + 2q\phi$$

where D is the deviance of the fit and, q is the number of estimable parameters and ϕ is the scale factor which is nominally 1 for binomial models, it was determined that neither the phase by diphone interaction PCV, nor the diphone interaction CV. was necessary to the model. It has been shown that AIC minimisation biases model selection towards more complicated models and it is therefore unlikely to remove significant terms[]. The AIC scores for three nested models are given in Table 1. Because the absence of the phase by vowel interaction term causes such a small increase in the AIC, it was decided to omit it from the final model under the assumption that any less liberal model selection procedure would not retain it.

	df	**AIC**
<base>	108	3070.621
-phase.vowel	90	3071.088
-phase.onset	78	3161.699

Table 1: AIC scores for 3 nested models.

Thus the final model that was selected for further examination was P + C + V + PC.

An examination of the individual subject coefficients for the P.C. term shows that minimum phase reconstruction has the most adverse affect on the perception of the alveolar consonants /d/ and /t/ evidently in favour of the other categories. Figures 1, 2, and 3 show a stimulus that gave 100% responses to either /t/ or /d/ for unaltered stimuli and 80% responses to either /b/ or /p/ for the minimum phase reconstruction. A possible suggestion for this would be the shift of low frequency energy towards the beginning of the signal which makes the spectrum at the onset more diffuse.

A multivariate analysis of variance on the individual subject coefficients for P anc PC showed no significant effects with a large degree of between subject variability. The power of this test is quite weak because of the small number of subjects — coefficients for PC absorb 5 degrees of freedom. However, within each subject, very large significant effects were found for PC ($\alpha < .01$ for four subjects, $\alpha < .1$ for two). This may indicate that while the minimum phase transformation has a significant effect on correct classification of stop consonant place of articulation, the effects are quite different accross listeners. For example, the coefficient for the PC term was largest and most significant for /d/ and /t/ except for one listener who showed a larger coefficient for /b/ and /p/.

5. LOGISTIC MODELING

In order to assess the importance of dynamic spectral information in theses stimuli, subjects' responses to unaltered stimuli were modeled by a logistic regression using mel ceptra. The fitted models were then used to make predictions on the minimum phase stimuli.

Two sets of logistic regressions were fit based on subjects' responses to the *unaltered* stimuli: one using mel cepstra from the full 23.75 ms signal. The other calculated mel cepstra from the first half hamming window weighted half as well as from the second half in order to exploit any dynamic spectral information present in the stimuli. In order to make the number of degrees of freedom equal, the dynamic model calculated 6 cepstra from each window while the static representation used 12 cepstra from the entire signal. Table 2 shows the percent modal agreement of the two models' predictions to both unaltered and minimum phase stimuli with subjects' responses. Despite the fact that the dynamic model uses fewer cepstral coefficients, it still makes more accurate predictions on subjects responses for unaltered stimuli. However there appears to be little difference between predictions made for minimum phase stimuli.

	unaltered	$x_{min}[n]$
dynamic	76.2%	65.9%
static	71.8%	65.3%

Table 2: Percent modal agreement of static and dynamic models with subjects' responses.

6. CONCLUSIONS AND DISCUSSION

The details of the results are complex and data from more listeners and more stimulus conditions will likely be necessary before any highly specific conclusions can be drawn with confidence. However, the present experiment provides reasonably strong evidence that change in the short time spectral structure of even very brief stimuli have important effects on the perception of place

of articulation in stop consonants. Further results will be presented at the meeting.

Research support by SSHRC.

7. REFERENCES

1. Stevens, K.N., and Blumstein, S.E. “Invariant cues for place of articulation in stop consonants,” *J. Acoust. Soc. Am.*, 64:1358–1368, 1978.
2. Blumstein, S.E., and Stevens, K.N. “Perceptual invariance and onset spectra for stop consonants in different vowel environments,” *J. Acoust. Soc. Am.*, 67: 648–662, 1980.
3. Kewley-Port, D. “Time-varying features as correlates of place of articulation in stop consonants,” *J. Acoust. Soc. Am.*, 73:322–335, 1983.
4. Nossair, Z.B., and Zahorian, S.A. “Dynamical spectral features as acoustic correlates for initial stop consonants,” *J. Acoust. Soc. Am.*, 89:2978–2991, 1993.
5. Lahiri, A., Gewirth, L., and Blumstein, S.E. “A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study,” *J. Acoust. Soc. Am.*, 76:391–404, 1994.
6. Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. “Statistical analysis of word-initial voiceless obstruents: Preliminary data,” *J. Acoust. Soc. Am.*, 84:115–123, 1988.
7. Tekieli, M. E., and Cullinan, W. L. “The perception of temporally segmented vowels and consonant-vowel syllables,” *J. Speech Hear. Res*., 22:103–121, 1979.
8. Smits, R., ten Bosch, L., and Collier, R “Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment,” *J. Acoust. Soc. Am.*, 100:3852–3864, 1996.
9. Oppenheim, A.V., and Schafer, R.W. *Discrete-Time Signal Processing*, Prentice Hall, Englewood Cliffs, 1989.
10. Nearey, T., and Assmann, P. “Modeling the role of inherent spectral change in vowel identification,” J. Acoust. Soc. Am., 80:1297–1308, 1986.
11. McCullagh, P., and Nelder, J.A. Generalized Linear Models, Chapman & Hall, London, 1989.
12. Venables, W.N., and Ripley, B.D. *Modern Applied Statistics with S-PLUS*, Springer, New York, 1997.