THE PERCEPTION OF SPECTRALLY REDUCED PREVOCALIC STOP CONSONANTS

Michael Kiefte
Department of Linguistics, University of Alberta, Canada

ABSTRACT

The present paper investigates the recoverability of detailed spectral features such as formant and burst peak frequencies of prevocalic stop consonants in noise excited channel vocoded speech for bandwidths ranging from 250–2000 Hz. It is shown that some formant frequency information is still recoverable up to 1000 Hz. This challenges the claim that listeners must necessarily rely primarily on temporal or global spectral features because such narrowly localized spectral properties are lost in noise vocoded speech. Listeners' categorization of these stimuli were modelled using empirically measured formant and burst peak frequencies as well as mel-frequency cepstral coefficients in order to evaluate the relative importance of each as correlates to place of articulation perception. It is found that despite a certain level of recoverability for detailed spectral features, spectral shape at burst and voicing onset are better correlates to listeners' perception of stop consonant place of articulation.

1. INTRODUCTION

In their work on simulating speech perception by recipients of cochlear implants with only a few processing channels, Shannon *et al.* [9], have shown that normal hearing listeners can be trained to identify stop consonant place of articulation from only four channels of amplitude modulated noise (hereafter referred to as noise vocoded speech). They conclude that speech perception can be achieved using primarily temporal cues.

However, it can be argued that this signal manipulation preserves gross spectral information including such features as global spectral tilt and "compactness" [13] of the burst onset as well as the change in spectral tilt at the onset of voicing [5] or alternatively, the spectro-/temporal envelope over the first few tens of milliseconds following oral release [4]—all of which have been postulated as important cues for stop consonant place of articulation. For example, in the case of a four channel implant simulation, we can assume that the energy in the lowest and highest frequency bands, as well as the relative difference between the two middle bands are roughly able to preserve spectral tilt; this is verified in Section 2.3. Shannon *et al.*'s results would lend support to the view that such grossly defined spectral shape features are robust cues in this context.

It has also been claimed that vocalic formant transitions show some relational invariance for place of articulation in stops (notably Sussman *et al.*'s locus equations [14]). Formant frequencies, which represent an example of a narrowly localized or "detailed" spectral feature (using Smit's terminology [12]), are assumed to be much more susceptible to degradation in frequency resolution such as that found in noise vocoded speech. This claim can be justified in a number of ways: for example, an 8th order LPC model will require at least a 9 point DFT for the coefficients to be recoverable from its Fourier transform. For a Nyquist cutoff at 4000 Hz, this entails that each Fourier coefficient covers a bandwidth of 500 Hz.

Does this mean that formant are not recoverable from noise vocoded speech in which each channel is broader than 500 Hz? If this were true, then listeners in Shannon *et al.*'s experiments [9] may be attending primarily to gross spectral shape cues in addition to any temporal information preserved by the signal processing. In the following section we determine how much formant frequency information is recoverable from noise vocoded speech with channel bandwidths between 500–2000 Hz. We compare this to the amount of variability introduced to global spectral shape measures by this type of spectral reduction.

In Section 3 we describe a perception experiment in which *untrained* listeners categorized place and voicing in noise vocoded prevocalic stop consonants in an attempt to evaluate the relative importance of spectral shape *vs.* detailed spectral cues such as formant and burst peak frequencies by explicitly modeling subjects' responses.

2. RECOVERABILITY OF GROSS AND DETAILED SPECTRAL CUES

2.1. Stimuli.

/CVk/ syllables were produced by twelve speakers of Western Canadian English (6 males and 6 females) in which /C/ was one of /b,d,g,p,t,k/ and where /V/ was one of /e,æ,ɒ,o/ since these vowels have been identified as the extreme points for F1 and F2 for this dialect [8]. In total, 6×4×12=288 syllables were used.

Syllables were low-pass filtered and digitized using a DT2821 AD/DA at either 10 or 12 bps and 16 kHz.

All syllables were subband decomposed via a polyphase-DFT realization of a uniform filter bank where the bandwidth of each channel is constant and the lowest and highest frequency channels are centered on the DC and Nyquist frequency respectively [1]. The analysis filters were derived from the polyphase decomposition of a high-order (134×the number of channels) linear phase FIR filter. After each channel was downsampled, they were replaced with signal correlated noise. Channels were then recombined by upsampling and refiltering. The result is a signal in which the amplitude envelope of each channel is preserved but the instantaneous frequency is randomized within each subband. This is similar to Shannon *et al.*'s stimuli with the exception that the analysis filters are linear phase and that each channel has equal bandwidth. Four different bandwidths were considered: 250, 500, 1000 and 2000 Hz.

2.2. Formant and Burst Peak Frequencies

			voicing onset						vowel steady state					
	burst peak		F1		F2		F3		F1		F2		F3	
BW	pct	rms	pct	rms	pct	rms	pct	rms	pct	rms	pct	rms	pct	rms
250 Hz	0.91	718	0.68	110	0.83	132	0.73	144	0.75	100	0.84	116	0.74	142
500 Hz	0.89	903	0.58*	144	0.73*	194	0.65	200	0.65*	122	0.74*	172	0.63*	195
1 kHz	0.91	844	0.56*	202	0.61**	237	*0.52***	257	0.58**	216	0.59**	236	0.58**	251
2 kHz	0.76**	1273	*0.51***	214	*0.52***	339	*0.52***	243	*0.52***	205	*0.5***	296	*0.53***	218

Table1. Probability of closer match between original and paired noise vocoded frame and randomly selected noise vocoded frame and rms errors between calculated noise vocoded and original values. Units are Hz.

With the aid of a wide band spectrogram and waveform, the onsets of voicing and vowel steady states were located for each unpocessed syllable using a procedure similar to Sussman *et al.* [14]. For each onset and steady state, a single glottal pulse was excised from the signal after downsampling to 8000 Hz, preempesized and weighted by a Kiaser window. The first three formants of each excised glottal pulse were measured empirically using either 8^{th} or 10^{th} order LPC. Candidate formant poles were rejected if they did not create an additional peak in the spectrum. The order of the LPC analysis was chosen based on which gave the better alignment to the formant peaks as observed on a spectrum of the frame as well as the wideband spectrogram.

A 10^{th} order LPC was calculated for each of the noise vocoded frames in the same way using the frame boundaries determined from the original signals. For each pole identified as a formant in the unprocessed frame, the pole in the noise vocoded frame nearest in Euclidean distance on the z-plane was selected and labelled as that formant.

Because this procedure is biased towards small rms errors in predicting the original formant frequencies from noise vocoded signals, each of the original formant poles was also compared to those of a randomly selected noise vocoded frame of the same type (either F2 onset or vowel steady state) and bandwidth. If the behavior of the LPC analysis is truly random for noise vocoded speech at a given analysis bandwidth, we expect that each undistorted formant will be nearer either to one pole of any randomly selected noise vocoded frame or to the frame that was actually prodeuced from the original signal with equal likelihood. Therefore we can compare this probability to a binomial distribution with $\pi=.5$ as a non-parametric test of randomness. However, we have no theoretical basis to determine the upper limit of this probability as a measure of LPC accuracy. Therefore the 250 Hz bandwidth processing condition was used as a baseline for comparison.

Table 1 shows the results of this non-parametric test for F1, 2, and 3 at the onset of voicing as well as for the vowel steady state. Significant differences between the 250 Hz bandwidth and all others are indicated by * for $\alpha \leq .05$ and by ** for $\alpha \leq .01$. Probabilities were also compared to a completely random null hypothesis—*i.e.*, $\pi=.5$—and non-significant differences are indicated by *italics* ($\alpha \leq .05$). In addition, rms values are given for the error in predicting the original formant frequency from the noise vocoded frame. As can be seen from the table, there is a drop in the performance of LPC between 250 and 500 Hz. However, LPC does shows significantly non-random behavior up to 1 kHz.

In addition to formants, burst peak frequencies for each stop were determined by locating the point of maximum spectral energy above 700 Hz from a Kaiser weighted segment containing only the burst (care was taken to not include the aspiration portion of voiceless stops). In contrast to the formant frequency estimation, burst peak frequencies are far more robust in noise vocoded speech; although rms increases with bandwidth, the statistical procedure used to test the randomness of the burst peak in noise vocoded speech does not show significant differences from 250 Hz up to 1 kHz bandwidth. The burst peak frequency appears to be well preserved in noise vocoded speech.

2.3. Mel-Frequency Cepstral Coefficients

We expect gross spectral properties such as global spectral tilt to be relatively well preserved in noise vocoded speech with only a few processing channels. To test this assumption, we followed a similar procedure as in Section 2.2 for mel-frequency cepstral coefficients (MFCCs).

Syllables were processed by a bank of forty triangular filters with equidistant center frequencies in the mel scale. The discrete cosine transform of the log amplitude for each frame was calculated [11]. The first basis function gives more weight to low frequencies and provides a good approximation to spectral tilt.

Cepstral coefficients were determined for two 25 ms frames: one starting at the onset of the burst and one at the onset of voicing. Randomization tests were performed and the results are given in Table 2.

Results show that the first three cepstral coefficients calculated at the release show significantly non-random behavior up to 2 kHz for $\alpha \leq .01$, as well as for the second and third coefficients at the onset of voicing. In addition, the first cepstral coefficient at the release burst, which roughly corresponds to global spectral tilt, shows no significant differences between 250 and 2000 Hz indicating that this property is well preserved in noise vocoded speech.

2.4. Discussion

We have shown that, although there is distortion in formant frequency estimation via LPC, some information is preserved up

BW	Burst onset			Onset of Voicing		
250 Hz	.72	.94	.92	.69	.86	.87
500 Hz	.80	.93	.90	.60*	.88	.90
1 kHz	.71	.77**	.65**	.39**	.78**	.59**
2 kHz	.66	.78**	.66**	.26**	.64**	.60**

Table 2. Probability of closer match between original and paired noise vocoded frame than randomly selected noise vocoded frame.

to 1 kHz noise vocoding bandwidth. In addition, the burst peak frequency is particularly robust in noise vocoded speech up to 2 kHz.

It has also been shown that low order mel-frequency cepstral coefficients—and in particular global spectral tilt at burst onset—are also well preserved.

While either may account for high performance levels for listeners of noise vocoded speech, we do not know if systematic misclassification of stop place of articulation is due to distortion in detailed spectral information such as formant or burst frequencies, or because of changes in global spectral properties such as gross spectral shape at burst onset or at the onset of voicing.

The next section describes a perceptual experiment designed to evaluate both these sets of cues via explicit modeling of listeners' responses to noise vocoded speech.

3. PERCEPTION EXPERIMENT

3.1. Subjects

Nine graduate and undergraduate students of Linguistics were paid as subjects in a speech perception experiment. None reported any hearing impairment and all were native speakers of Western Canadian English.

3.2. Procedure

Stimuli from the twelve speakers described in Section 2.1 were presented to subjects for classification of the syllable initial stop. Only 500 and 1000 Hz noise vocoded stimuli were included in the set of spectrally distorted stimuli. In addition to these, the original stimuli were presented after having been decomposed/resynthesized *without* substitution with signal correlated noise in order to evaluate the effects (if any) of the filter bank reconstruction (which introduces some noise) as a baseline. In total, 864 stimuli were presented to listeners in two sessions of approximately 25 minutes each.

All stimuli were completely randomized and presented to listeners who received *no prior training.* This is in contrast to procedure described by Shannon *et al.* [9] in which listeners received extensive training. Subjects were asked to indicate by clicking on the appropriate button on a computer screen which consonant they thought began each syllable.

Stimuli were played through a Gina AD/DA at 16bps and 44.1 kHz on a PC. Subjects heard stimuli in a sound treated room at a comfortable listening level.

	reconstructed			500 Hz			1000 Hz		
	lab	alv	vel	lab	alv	vel	lab	alv	vel
lab	.65	.01	.00	.60	.04	.02	.58	.04	.04
alv	.02	.64	.01	.06	.58	.03	.12	.53	.02
vel	.00	.02	.65	.02	.06	.59	.11	.16	.40

Table 3. Probability of correct classification of place of articulation.

3.3. Results

The mean probability of correct classification of place of articulation over the nine listeners was .97, .89 and .76 for the reconstructed, 500 Hz noise vocoded and 1000 Hz noise vocoded syllables respectively as compared to .33 for chance. McNemar tests [3] showed that there were significant differences in performance between correct classification for the reconstructed stimuli and the 500 bandwidth noise vocoded stimuli as well as between 500 and 1000 Hz bandwidth stimuli for all subjects at the $\alpha \leq .001$ level of significance.

Responses to place of articulation (voicing was discarded for the purposes of this analysis), was fit with a generalized linear model in which the multinomial frequency data was treated as a Poisson log-linear process conditional on the total number of responses per stimulus [6]. Although this was done using an iterative weighted least squares regression, this is equivalent to a single layer neural network using maximum conditional likelihood fitting used by Smits *et al.* [12].

Initially, four simple fixed effects were considered: actual place of articulation, voicing of the prevocalic stop, identity of the vowel and whether the stimulus was channel reconstructed, or was noise vocoded at either 500 or 1000 Hz bandwidth. The optimal model was selected by backwards stepwise minimization of the Akaike information criterion (AIC) which is based on the residual deviance plus two times the number of estimated parameters [6]. A separate model was fit for each subject and the deviances from each subject was summed along with the total number of parameters to determine the AIC.

The only processing condition interaction effect included by the optimized model was place×processing. Although it is known that minimization of the AIC biases model selection towards the larger of nested models, it is unlikely to exclude interactions that are significant [10].

The coefficients (weights) for the place×processing interaction show that responses to alveolar and velar place of articulation drop with increased spectral distortion. When coefficients from individuals are treated as random variables [7], this difference is significant at the $\alpha \leq .05$ level only for alveolar responses to alveolar stops between 500 and 1000 Hz bandwidth noise vocoded speech.

Table 3 gives the probabilities for correct identification of stop place of articulation for each processing condition. The actual place of articulation is given along the left and response labels are given along the top of the table. It shows that incorrect labial responses increase with increased spectral distortion as do incorrect alveolar responses to velars. Interestingly, this seems to parallel the behavior observed in the perception of burstless stops

Factors	*e.d.f.*	AIC	Dev.
Burst and formant frequencies	28	13101	13045
burst and onset cepstra (1–3)	20	12784	12745
burst cepstra (1–3) and form.	26	12447	12395
burst and onset cepstra (1–4)	26	12297	12245

Table 4. Analysis of deviance for four models including formant/burst peak frequency and MFCCs.

[12]. Why this should be the case is difficult to explain, since in section 2.2, we show that burst peak frequency is a robust cue in spectrally reduced stimuli. What it perhaps indicates is that while spectral peak is well preserved, it is not a good correlate for place of articulation perception.

Data were fit with a generalized log-linear regression using estimated spectral parameters as continuous covariates. The first model considered consisted of burst peak frequency and formant frequencies measured at the onset of voicing and at the vowel steady state. Because locus equations have been shown to provide some separability between different places of articulation [14], interactions between onset formant frequencies and the corresponding formant frequencies at vowel steady state were also included (*e.g.*, $F2_{ons}$ by $F2_{vow}$ interactions). In addition, because it is known that burst peak frequencies are dependent on vowel context (most notably for velars [2]), all burst peak by onset formant frequencies were also included (which produced a lower deviance than burst peak by vowel formant interactions). Table 4 shows the total number of estimated parameters "*e.d.f.*", the AIC and the residual deviance for this and the following models.

The second model considered used the first three MFCCs measured at the burst onset as well as at the onset of voicing. Based on Lahiri *et al.*'s suggestion that the relative change in spectral tilt between these two frames may be an important correlate to place perception [5], all interactions of same index coefficients in these two frames were also included.

Despite having fewer estimated coefficients, this model has a lower residual deviance than the first, suggesting that global spectral measures provide a better fit to the data.

The third model included the first three cepstral coefficients measured at the burst as well as formant frequency covariates used in the first model. Again this model estimates fewer parameters than the first but has a lower residual deviance. Although it contains more parameters than the second model, the AIC is lower.

To generate a model with as many estimated parameters as the third using only cepstral coefficients, a forth model was fit using the first *four* MFCCs measured at the onset of release burst and at the onset of voicing along with all pairwise interactions. The model has the lowest residual deviance and although it contains more parameters, has a lower AIC than the second.

Although model selection by minimization of AIC is biased towards larger models the magnitude of the difference between the second and fourth models appears large relative to the increase in *e.d.f.* However, a conclusion cannot be made until a less biased criterion is used.

4. DISCUSSION

Absent from the analysis in Section 3.4 are other potentially important cues such as burst peak amplitude, and burst duration. In addition, formant measurements were not made during the aspiration portions of voiceless stops, which are much more difficult to extract, but may provide a better fit to listeners' responses. In addition a more vigorous comparison of models must be made using an unbiased procedure such as bootstrap model selection [10].

However, it would appear that, overall, MFCCs are better correlates to stop consonant place of identification than detailed formant and burst peak frequency estimates. This may be related to the difficulty in calculating the latter, but it could also be related to the robustness of global spectral shape measures in noise vocoded speech as indicated by the randomization tests in Section 2.3.

ACKNOWLEDGMENTS

This work was supported by SSHRC.

REFERENCES

[1] Akansu, A. N., and Haddad, R. A. 1992. *Multiresolution Signal Decomposition.* Boston, Academic Press.

[2] Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. 1952. Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24. 597–606.

[3] Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions.* Second edition. New York: Wiley.

[4] Kewley-Port, D. 1983. Time-varying features as correlates of place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 73. 322–335.

[5] Lahiri, A., Gewirth L., and Blumstein, S. E. 1984. 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.

[6] McCullagh, P., and Nelder, J. A. 1989. *Generalized Linear Models.* Second edition. London, Chapman and Hall.

[7] Nearey, T. M. 1997. Speech perception as pattern recognition. *J. Acoust. Soc. Am.*, 101. 3241–3254.

[8] Nearey, T. M., and Assmann, P. 1986. Modelling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.*, 80. 1297–1308.

[9] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. 1995. Speech recognition with primarily temporal cues. *Science,* 270. 303–304.

[10] Shao, J. 1996. Bootstrap model selection. *J. Am. Statist. Assoc.*, 91. 655–666.

[11] Slaney, M. 1994. Auditory toolbox. *Apple Technical Report* 45.

[12] Smits, R., ten Bosch, L., and Collier, R. 1996. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *J. Acoust. Soc. Am.*, 100. 3852–3864.

[13] Stevens, K. N., and Blumstein, S. E. 1978. Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 64. 1358–1368.

[14] Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. 1991. An investigation of locus equations as a source of relational invariance for stop place of articulation. *J. Acoust. Soc. Am.*, 90. 1309–1325.