PERCEPTION OF SPECTRALLY COMPRESSED SPEECH

R.R.HURTIG

Dept. of Speech Pathology & Audiology
University of Iowa, Iowa City, Iowa, USA

ABSTRACT

The effect of spectral compression and frequency transposition on the perception of vowels and simple sentences is examined. A computational algorithm which can accomplish spectral compression and frequency transposition is presented. Analysis of confusion data suggests that spectral shape rather than absolute formant frequency differences is critical to vowel identification. The limits of frequency transposition appear to be determined by critical bark differences.

1.INTRODUCTION

The perception of speech, and in particular vowels, is in many respects conditioned by both static and dynamic cues in the speech spectrum. The centrality of the first and second formants in vowel identification is generally accepted. At one level it has been suggested that the detection of spectral prominences (formants) provides the necessary and sufficient acoustic cues to the vocal tract configuration which is associated with a particular vowel production. This position must be tempered somewhat by the "speaker normalization effect" and the range of formant values for speakers of very different ages. Recent accounts of vowel perception [5,6] suggest that an extension of the "center of gravity" hypothesis [1] may provide the basis for the detection of vowels based on the notion that a given vowel can be represented in terms of the extent to which formant frequency differences are greater than or less than 3 Bark. Thus it may be a categorical difference rather than an absolute difference in formant frequencies that specifies a particular vowel.

This leads to a prediction that only a subset of the signal processing schemes should preserve vowel identification and that other transformations should lead to a decrement in speech perception. Specifically, any processing of the speech spectrum which preserves the Bark difference values distinguishing vowels will result in adequate speech perception. By contrast any transformation which alters the distinctions marked by the Bark difference values of the natural stimuli should render the speech unintelligible and perhaps alter the signal sufficiently to make it lose its speech quality.

Such an interpretation holds that within limits it is the spectral shape (i.e., slope) which is the invariant cue to vowel identity. Thus the acoustic cues are not necessarily formant values associated with normal vocal tract configurations but are rather values which can be associated with a scaled vocal tract. To test this general hypothesis an algorithm for processing the speech spectrum was selected which would maintain the spectral shape but which would allow a scaling of the spectrum (frequency compression) and a transposition of the spectrum (frequency shift).

2.PROCEDURE

2.1 TFT-Algorithm

A computational algorithm [2] was used to achieve spectral compression and frequency transposition of natural speech samples. The TFT (Time-to-frequency-to-time domain) algorithm performs its manipulations of the signal in the frequency domain. The algorithm operates on successive windows of n-samples. First an n-point FFT is computed. The resultant complex array is padded with the spectrum of a Hamming window. An IFFT is then computed on the padded array. The resultant real array is then trimmed to the length of the original input window. The size of the pad relative to the size of the input window determines the degree of compression to be achieved. The placement of the pad relative to the complex array determines the degree of frequency transposition. For a pad of a length equal to the input window a frequency compression to 50% of the original spectrum is achieved; a pad three times the length of the input window yields a compression to 25% of the original. Positioning the pad following the complex array yields compression with no frequency transposition of the resultant spectrum. By contrast a leading pad would achieve both compression and an upward shift in the frequency of the resultant spectrum. A partitioning of the pad into a leading and following component can yield varying degrees of frequency transposition based on the proportion of the pad in the leading position.

2.2 Stimuli

Multiple tokens of natural vowels [i,ɪ,ɛ,æ,a,ʌ,ɔ,o,ʊ,u] in the hVd context as well as sentences from short narrative passages were digitized and subjected to the TFT algorithm. These stimuli were compressed to 50% or 25% of the original and were also frequency transposed in 5 steps from 150 Hz to 2500 Hz.

2.3 Presentation and testing of vowel tokens.

Listeners (N=7) were given the opportunity to explore a set of vowels at a given compression and frequency transposition using a listener paced exploratory learning task in which the listener is free to select the specific token to be played out. The listeners explored a given set of stimuli for a minimum of fifteen minutes after which the computer presented the tokens in a random order in a closed set recognition paradigm.

2.4 Presentation and testing of the sentence tokens.

The sentence tokens were presented in either a random order or in short connected discourses. Listeners (N=9) were required to write down as much of each sentence as they could. In addition some listeners were presented sentences with varying degrees of frequency transposition for open set recognition and judgments of speech quality and intelligibility.

3. RESULTS

3.1 Vowels.

The correct identification of vowels was 44% in the 50% compression condition and 81% in the 50% compression with a 150 Hz transposition condition. These scores represent performance which is significantly better than chance performance. All listeners showed a marked improvement in the 150 Hz transposition condition. An Information Transfer Analysis [3] reveals the relative information transferred for the first three formants, for each of the conditions. It should be noted that in both conditions there were comparable transfer rates for each of the format cues.

Table 1.
Percent Relative
Information Transferred

	F1	F2	F3
50%	26	33	27
50%(+150Hz)	73	73	74

The best subject's performance was 64% in the 50% compression condition and 86% in the condition with 150 Hz transposition. Table 2 indicates near perfect transfer of formant information in the transposed condition.

Table 2.
Best Subject Percent
Relative Information Transferred

	F1	F2	F3
50%	53	51	44
50%(+150Hz)	92	92	93

3.2 Sentences.

Tracking was assessed in terms of the percentage of words correctly identified. The tracking of speech compressed to 50% is fairly easy to do and requires effectively no listening experience. Tracking 50% compressed sentences with a 150 Hz frequency transposition was better than tracking untransposed sentences. Likewise tracking of 50% compressed speech was better than the tracking of 25% compressed speech. As expected, providing a context improves tracking.

Table 3.
Percent Correct Tracking

	No Context	Context
50%	87	99
50%(+150Hz)	97	100
25%(+150Hz)	32	67

The subjective quality of the resultant frequency compressed speech varies with the degree of frequency transposition. The most natural sounding compressed tokens are those with a minimal amount of frequency transposition (from about 150-500 Hz). As the frequency transposition exceeds 500 Hz the intelligibility and naturalness of the tokens deteriorates.

4.0 DISCUSSION

The ability to identify vowels and to track running speech which has been frequency compressed argues that speech perception is the consequence of a process which is sensitive to spectral shape rather than specific spectral prominences which can be associated with formants produced by natural vocal tracts. The limitation on this ability appears to be based on the critical bandwidth of the auditory system's filters [7]. A linear transposition of the spectrum in the frequency domain will result in a greater probability of formant peaks falling within a single critical band. At the extreme the signal loses any resemblance to speech and has a cricket-like quality. This effect is easily predicted from the Bark difference scores which result from linear frequency transpositions. Using the formant values from the TI data base [5] one can calculate the bark difference values for both natural vowels as well as those with 50% and 25% compression and various amounts of linear frequency transposition. (As Figure 1 illustrates the natural distinctions are basically maintained with transposition of less than 650 Hz.)

These data are consistent with the results obtained with frequency transposition hearing aids which fail to yield results better than those obtained with simple filter circuits. The tinny percept is perhaps similar to unpleasant percept reported by cochlear implant users. This may be due to typical electrode insertion which stimulates fibers associated with higher frequencies and correspondingly internal filters with wider critical bandwidths.

A: F2-F1

	TI	50%	50% 150	50% 350	50% 650	50% 1250	50% 2500	25%
heed							+	
hid						+	+	
head						+	+	
had					+	+	+	+
herd				+	+	+	+	+
huhd		+	+	+	+	+	+	+
hahd	+	+	+	+	+	+	+	+
hod	+	+	+	+	+	+	+	+
hood				+	+	+	+	+
huwd				+	+	+		

B: F3-F2

	TI	50%	50% 150	50% 350	50% 650	50% 1250	50% 2500	25%
heed	+	+	+	+	+	+	+	+
hid	+	+	+	+	+	+	+	+
head	+	+	+	+	+	+	+	+
had	+	+	+	+	+	+	+	+
herd	+	+	+	+	+	+	+	+
huhd					+	+	+	+
hahd					+	+	+	
hod						+	+	
hood				+	+	+	+	+
huwd		+	+	+	+	+	+	

Figure 1.
Bark Difference values for vowels in the TI Data Base and for compressed and frequency transposed versions. (+ indicates a difference of <3 Bark)

5. REFERENCES

[1] Chistovich, L.A. & Lublinskaya, V.V. (1979), "The 'center of gravity' effect in vowel spectra and critical distance between formants: Psychoacoustical study of the perception of vowel-like stimuli.", Hear.Res., Vol.1, 185-195.

[2] Hurtig, R.R. (1989), "TFT, An algorithm for the spectral compression of natural speech signals.", J.Acoust.Soc.Am., Suppl.1, Vol. 85, S44.

[3] Nabelek, A.K. & Letowski, T.R. (1985) "Vowel confusions of hearing-impaired listeners under reverberant and nonreverberant conditions.", JSHD, Vol. 50, 126-131.

[4] Peterson, G.E. & Barney, H.L. (1952), "Control methods used in the study of vowels.", J.Acoust.Soc.Am., Vol. 24, 175-184.

[5] Syrdal, A.K. (1985), "Aspects of a model of the auditory representation of American English Vowels.", Speech Commun., Vol. 4, 121-125.

[6] Syrdal, A.K & Gopal, H.S. (1986), "A perceptual model of vowel recognition based on the auditory representation of American English vowels.", J.Acoust.Soc.Am., Vol.79(4), 1086-1100.

[7] Zwicker, E. & Terhardt, E. (1980), "Analytic expressions for critical-band rate and critical bandwidth as a function of frequency.", J.Acoust.Soc.Am., Vol.68(5),1523-1525.