

MODELING THE PERCEPTION OF FREQUENCY-SHIFTED VOWELS

Peter F. Assmann¹, Terrance M. Nearey²,
and Jack M. Scott³

^{1,3} School of Human Development
University of Texas at Dallas, Richardson TX 75083
assmann@utdallas.edu

² Department of Linguistics
University of Alberta, Edmonton AB T6G 2E7
t.nearey@ualberta.ca

ABSTRACT

A significant fact about speech perception is that intelligibility is preserved when the spectrum is shifted up or down along the frequency scale, across a fairly wide range. To study the relationship between fundamental frequency (F_0) and spectrum envelope shifts in vowel perception, we used a high-quality vocoder (STRAIGHT) to process a set of vowels spoken by 3 adult males in /hVd/ context. Identification accuracy dropped by about 30% when the spectrum envelope was scaled upwards by a factor of 2.0, and in a separate condition, by about 50% when F_0 was raised by 2 octaves. However, when spectrum envelope and F_0 were both increased at the same time, identification accuracy showed a marked improvement, compared to conditions where each cue was manipulated separately. The synergy between formant frequency and F_0 was predicted by a model which accounts for the intelligibility of frequency-shifted vowels in terms of learned relationships between measured values of F_0 and formant frequencies. A second model, based on auditory excitation patterns, predicted the main effects of F_0 and spectrum envelope, but did not predict the pattern of interaction.

1. INTRODUCTION

The ability to understand frequency-shifted speech is a prerequisite for everyday speech communication, because listeners must adapt to variations in fundamental frequency (F_0) and formant frequencies associated with size differences in the larynx and vocal tract across talkers (Nearey, 1989; 1998). This ability must be explained by models of speech perception, and research on this topic may provide insights into two problems faced by hearing-impaired listeners. First, present-day cochlear implant electrode arrays cannot be inserted completely into the cochlea, and provide electrical stimulation only to the basal portion (Fu and Shannon, 1999). Implant users may need to accommodate to the re-mapping of the frequency spectrum provided by the device. Second, frequency shifts are used in frequency-transposing hearing aids which attempt to restore speech intelligibility for impaired listeners by shifting the spectrum into the region of better hearing (McDermott et al., 1999). Frequency lowering provides improved speech

recognition for some hearing-impaired listeners, especially after extended exposure, but the limited extent of its benefit warrants further study. To study the effects of upward shifts in formant frequency and F_0 on intelligibility, we used a source-filter vocoder (Kawahara, 1997) to process a sample of vowels in /hVd/ words.

2. EXPERIMENT

The experiment was designed to measure the relative contributions of *spectrum envelope* and *fundamental frequency* to frequency-shifted speech. The results showed that vowel identification accuracy was reduced when the spectrum envelope was shifted upward by a factor of 2.0; accuracy was also lower when F_0 was raised by two octaves. However, when both shifts were applied together (producing a voice similar to that of a small child) performance *improved*, reaching the level of accuracy expected for vowels spoken by young children.

2.1. Stimuli

Eleven vowels in /hVd/ words (*heed, hid, hayed, head, had, hud, hawed, hoed, hood, who'd, herd*) spoken by 3 adult males were selected from a larger sample of vowels recorded from 10 men, 10 women, and 30 children, ages 3, 5, and 7 years from the North Texas region (Assmann & Katz, 2000; Katz & Assmann, 2001). From these original signals, 15 synthesized versions of each vowel were produced using the STRAIGHT program. These included 5 levels of spectrum envelope (or formant frequency, denoted **FF**) scale factor, combined with 3 levels of F_0 scale factor:

FF scale factor = 1.0, 1.25, 1.5, 1.75, 2.0

F_0 scale factor = 1.0, 2.0, 4.0

Vowel=/i/, /ɪ/, /e/, /ɛ/, /æ/, /ʌ/, /ɜ/, /ɑ/, /o/, /ʊ/, /u/

The scale factors were chosen on the basis of regression analyses of vowel formant frequencies and F_0 , combined with pilot listening. Adult male vowels were used in this study because the largest frequency shifts produced sounds like the voices of very small children; intermediate shifts were heard as female voices.

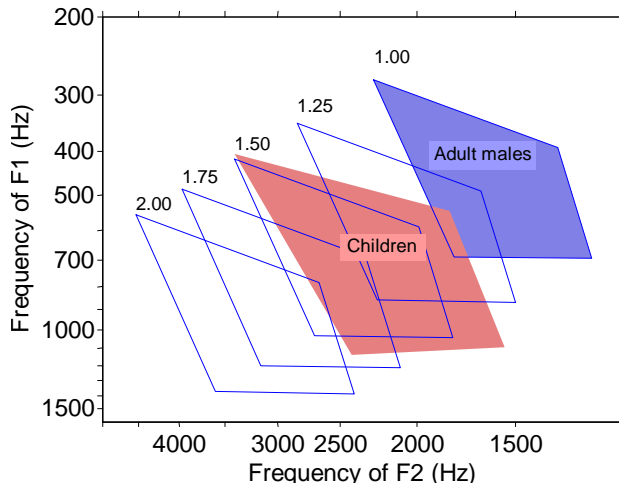


Fig. 1. Effects of spectrum envelope shifts

Figure 1 shows mean first and second formant frequencies, F1 and F2, from 10 adult males and 30 children (ages 3,5,7) in our vowel database for the corner vowels, /i/, /æ/, /ɑ/, and /u/ (shaded regions). The unshaded quadrilaterals indicate the effects on the male vowel space of the spectrum envelope shift factors used in the experiment. The two most extreme shifts in spectrum envelope, 1.75 and 2.0, generated vowels with formant frequencies that lie outside the range found in the youngest children in our database (3-year olds). The most extreme shift in F_0 (4.0) also produced vowels with higher pitches than those of the youngest children. We were particularly interested to see whether vowel identification declines when the shifts generate vowels whose formant frequencies lie outside the natural human range.

2.2. Listeners

Eleven young adult listeners with normal hearing were recruited from the undergraduate Psychology subject pool at the University of Texas at Dallas, and received partial course credit for their participation. Ten of the listeners were native speakers of American English from the North Texas region. They were required to complete a dialect questionnaire and a hearing screen (pure tone thresholds better than 25 dB HL at octave frequencies between 0.25 and 8.0 kHz).

2.3. Procedure

The experimental procedure was similar to that described in Assmann and Katz (2000). The stimuli were /hVd/ words presented monaurally over headphones in a double-walled sound booth at a mean level of 68 dB SPL (A). Listeners identified the vowels using an 11-category response box drawn on the computer screen. Prior to the experiment, listeners completed practice sessions until they reached a score of 85% or better on a set of /hVd/ words spoken by a different talker. In the main experiment they heard 495 stimuli (11 vowels x 3 talkers x 5 spectrum envelope shifts x 3 F_0 shifts). All conditions, vowels, and talkers were randomly interspersed.

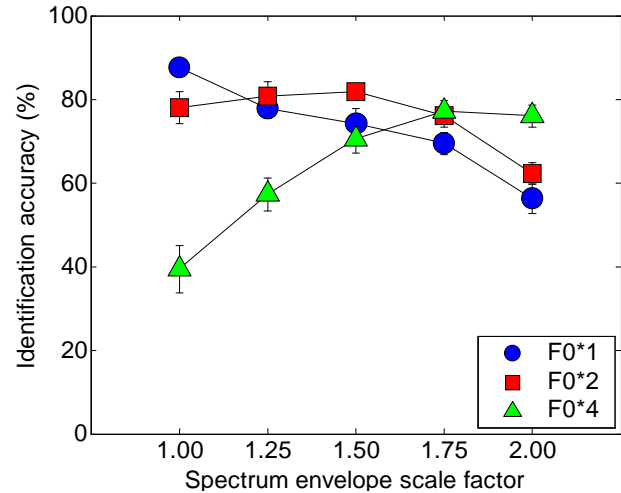


Fig. 2. Identification accuracy by listeners

2.4. Results and discussion

The results are shown in Figure 2. If the intelligibility of frequency-shifted speech is determined by listeners' sensitivity to patterns of statistical variation in natural speech, then vowels whose F_0 and formant frequencies lie outside the range of natural voices should be less accurately identified. This prediction was confirmed: extreme shifts in either F_0 or formant frequencies led to reduced vowel identification accuracy. However, when upward shifts in F_0 and formant pattern were applied simultaneously, identification accuracy showed a significant improvement.

One explanation is that *learned relationships between F_0 and formant pattern* are responsible for the effects of frequency shifts on vowel identification. Three aspects of the data lend support to a perceptual learning account. First, identification accuracy is lower for vowels whose formant frequencies lie outside the range of natural voices. Second, intelligibility is reduced when the formants and F_0 are *mismatched* (i.e. the formants are shifted into the region appropriate for a child, but the F_0 is fixed in the adult male range, or *vice versa*). Third, intelligibility is higher when F_0 and spectrum envelope are shifted together, presumably because this manipulation preserves the natural relationships between the frequencies of the formants and F_0 . Two pattern recognition models were implemented to simulate the perceptual data. Both models incorporate supervised learning algorithms, and were trained using a large database of adults' and children's vowels, representing a wide range of F_0 's and formant frequencies. Further support for the perceptual learning account would be provided if the statistical covariation between F_0 and formant frequencies in the training database leads the models to predict the interaction found in listener's identification responses.

3. THE MODELS

The working hypothesis is that the skills required for recognizing frequency-shifted vowels are the same as those

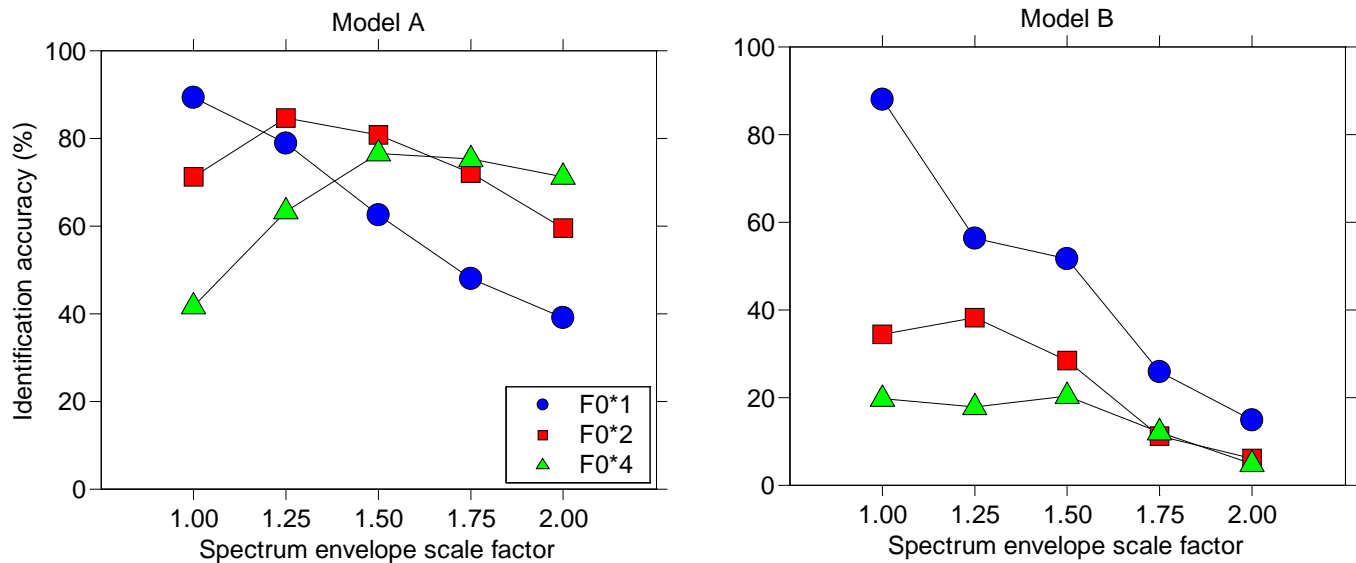


Figure 3: Model predictions for Experiment 1

used to recognize vowels spoken by talkers with different vocal tract sizes. Thus, the reduction in vowel identification accuracy may reflect a breakdown of listeners' abilities to adjust to acoustic patterns outside their range of experience with human voices. When frequency shifts produce vowels whose formant patterns (or spectral shapes) fail to match listeners' expectations, their phonetic quality becomes distorted and confusion errors are made. Two models of vowel classification were developed to test this hypothesis. The models consist of two stages, a training stage and a test stage. In the training stage each model extracts statistical regularities from a large sample of vowels to classify the test data. The parameters of the model are fixed once the training is complete.

4. MODEL A

Model A is an implementation of the pattern recognition model described by Hillenbrand and Nearey (1999). The input is a "dual-target" representation (Nearey and Assmann, 1986) with 8 parameters: mean F_0 , formant frequencies F_1 , F_2 , and F_3 sampled at the 20% and 80% points in the vowel, and vowel duration. The frequency measures are expressed in log units. Linear discriminant function analysis is used to generate *a posteriori* probabilities of group membership for each test vowel. In previous studies, this model has generated accurate predictions of vowel identification by listeners, including mean accuracy and the pattern of confusion errors (Hillenbrand and Nearey, 1999). For the purpose of the present study, a key assumption of the model is that listeners have internalized knowledge of the relationship between F_0 and formant frequencies in natural speech, i.e. that higher formant frequencies are accompanied by higher F_0 and vice versa. The training data for the model was a set of 2750 vowels (examples of each of the 11 vowels from 50 talkers, including 10 males, 10 females, and 30 children).

5. MODEL B

Model B is an implementation of a *whole-spectrum* model (e.g. Bakkum, Plomp and Pols, 1995). The model had three stages. First, the signal was analyzed using a bank of gammatone filters

(Patterson et al., 1992) with center frequencies between 0.1 and 5.08 kHz, equally spaced on a scale of ERB-rate (Moore and Glasberg, 1983). The output of each filter was analyzed by a sliding rectangular window, and the rms energy in the window was computed in eight frames of equal duration, chosen by dividing the total vowel duration into equal intervals. In the second stage, the excitation patterns were analyzed using a linear two-layer associative neural network with 1024 input units (128 filter channels x 8 time frames) and 11 output units (11 vowel categories). The network generates a set of output activations corresponding to each of the available response categories. In the third stage, input activations were mapped onto output activations via a set of connection strengths or weights, using a linear mapping function. Output activations were converted to predicted probabilities using a version of the Luce choice model. A related model was previously used to predict listeners' identification of concurrent vowels (Assmann, 1996). Although Model B does not provide an explicit representation of F_0 and formant frequencies, these parameters can be derived from an analysis of the excitation patterns (de Cheveigné and Kawahara, 1999).

6. MODEL EVALUATION

The predictions of the two models are shown in Figure 3 (compare with listeners' data in Figure 2). Both models predict the best performance for the *unshifted* condition, consistent with listeners. Moreover, both models predict the main effects, i.e. the drop in accuracy with upward shifts in F_0 and spectrum envelope, although Model B predicts larger declines with frequency shifts than shown by listeners. Model A predicts the interaction between F_0 and spectrum envelope, i.e. improved performance when F_0 and spectrum envelope are *both* shifted upward. Although there are discrepancies between the predicted and observed means for both models, it is clear that Model A comes closer to predicting the observed interaction between F_0 and spectrum envelope shifts in listeners' identification responses. However, the poorer performance of Model B may be linked to parameter settings and other implementation details that need further investigation.

7. DISCUSSION

The experiment showed that vowel identification accuracy was reduced when upward frequency shifts in the spectrum envelope were introduced. Identification also declined when F_0 was increased. However, identification accuracy was partially restored when upward shifts in both parameters were provided simultaneously. This outcome is consistent with the idea that learned relationships between formant frequencies and F_0 are important in vowel identification.

Two models of vowel identification were implemented to predict the identification responses of listeners. Both models predicted the decline in identification accuracy as a function of upward shifts in F_0 and spectrum envelope. However, only Model A predicted an improvement in performance when F_0 and spectrum envelope shifts were applied simultaneously. The correspondence between observed and predicted identification provides further support for the idea that the effects of frequency shifts in vowel perception can be explained by processes of perceptual learning through exposure to statistical variation in natural speech. We are currently investigating possible reasons while model B failed to predict this interaction. Model A incorporates an explicit representation of F_0 and formant frequencies, while model B uses a high-dimensional spectral representation derived from models of auditory masking and frequency selectivity. In principle, the formant frequencies and F_0 can be derived from the pattern of peaks in the excitation pattern (de Cheveigné and Kawahara, 1999) and hence the shortcomings of Model B are more likely to be related to the nature of the training procedure or the type of learning algorithm employed.

Overall, the results are consistent with the hypothesis that the perceptual effects of frequency shifts depend on learned relationships between F_0 and formant frequencies in natural speech. The experimental and modeling results provide evidence that the perception of frequency-shifted speech is constrained in important ways by listeners' sensitivity to statistical regularities in natural speech.

8. REFERENCES

- [1] Assmann, P.F. (1996). Modeling the perception of concurrent vowels: Role of formant transitions. *J. Acoust. Soc. Am.* 100: 1141-1152.
- [2] Assmann, P.F. & Katz, W.F. (2000). Time-varying spectral change in the vowels of children and adults. *J. Acoust. Soc. Am.* 108: 1856-1866.
- [3] Bakkum, M.J., Plomp, R., & Pols, L.C.W. (1995). Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children. *J. Acoust. Soc. Am.* 98: 745-762.
- [4] de Cheveigné, A. and Kawahara, H. (1999). Missing data model of vowel perception. *J. Acoust. Soc. Am.* 105, 3497-3508.
- [5] Fu, Q.-J. & Shannon, R.V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *J. Acoust. Soc. Am.* 105: 1889-1900.
- [6] Hillenbrand J.M., Nearey T.M. (1999). Identification of resynthesized /hVd/ utterances: effects of formant contour. *J. Acoust. Soc. Am.* 105: 3509-3523.
- [7] Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. *Proceedings of the ICASSP*, pp. 1303-1306.
- [8] Katz, W.F. & Assmann, P.F. (2001). Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. *J. Phonetics* 29: 23-51.
- [9] Moore, B.C.J., & Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* 74, 750-753.
- [10] McDermott, H.J., Dorkos, V.P., Dean, M.R., & Ching, T.Y.C. (1999). Improvements in speech perception with use of the AVR TranSonic frequency-transposing hearing aid. *J. Speech, Lang., & Hear. Res.* 42: 1323-1335.
- [11] Nearey, T.M. (1989). Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 85: 2088-2113.
- [12] Nearey, T.M. (1998). Selection of a tonotopic scale for vowels. In *Proceedings of the 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, (pp. 2001-2002). Seattle, WA: American Institute of Physics.
- [13] Nearey, T.M. & Assmann, P.F. (1986). Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.* 80: 1297-1308.
- [14] Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M.H. (1992) Complex sounds and auditory images. In *Auditory Physiology and Perception* (Y. Cazals, L. Demany, & K. Horner, editors), pp. 429-446. Oxford: Pergamon Press..