METHODS & DESIGNS

Generation of speech continua through monaural fusion

DALE C. STEVENSON, JOHN T. HOGAN, and ANTON J. ROZSYPAL
University of Alberta, Edmonton, Alberta, Canada

A procedure based on monaural fusion has been developed to construct acoustic continua between natural speech sounds, to be used in studies of speech perception. Two speech stimuli of similar temporal structure and different spectral composition are precisely aligned in time and presented simultaneously to the listener. By mixing both stimulus components in varying intensity ratios, a transition from one component to the other can be achieved. Such stimulus continua have several advantages over the synthetic continua commonly used in studies of categorical perception and related phenomena: They are based on real speech stimuli; the endpoint stimuli are unambiguous; and the stimuli are characterized by a well-defined physical variable, the relative intensity of the two components.

Traditionally, stimuli for recognition and discrimination experiments in speech research have been generated synthetically. Two endpoint stimuli representing the prototypes for the response categories are synthesized, along with a series of intermediate stimuli whose parameters are interpolated between those of the two prototypes, thus creating a continuum.

This article describes an alternative technique, monaural fusion, for constructing speech continua (Stevenson, 1979). The monaural fusion technique creates the stimulus continuum by aligning two speech stimuli of similar temporal structure but different spectral composition and presenting them simultaneously to the subject. The two stimulus components ordinarily fuse to produce a single percept. When the relative intensities of the two components are varied, a physical continuum that is categorically perceived is created. For instance, for syllable-initial stop consonants /b/ and /d/, the relative intensity continuum yields identification and discrimination results that are essentially equivalent to those obtained with traditional synthetic continua, obtained by interpolating formant frequencies (Repp, 1981; Stevenson, 1979).

Applicability, advantages, and limitations of this technique are discussed in the last section of this paper.

STIMULUS GENERATION

All signal preparation and presentation was carried out digitally with a system program written in assembly language and designed for the DEC PDP-12 minicomputer (Stevenson & Stephens, 1978). In this paper, emphasis is given to the method of generating the stimuli, rather than to the hardware and software, since these may vary between installations. Requirements on the frequency

characteristics and signal-to-noise ratio of the whole signal-processing channel must be specified by the experimenter and can be satisfied in the first place by choosing an adequate sampling rate and number of quantization levels at the speech digitization and desampling stages (Rozsypal, 1976).

The stimulus construction procedure is described in terms of stimuli /bæ/-/dæ/. Figure 1 shows a schematic spectrogram of an ambiguous stimulus, in which the superimposed formant trajectories of the two signals are shown as hatched bars for /bæ/ and as dotted bars for /dæ/. As can be seen, in this stimulus pair the temporal patterns, that is, the overall duration and the duration of the initial formant transitions, are matched. For all three formants in both stimuli, the formant frequencies asymptote toward the same value. The two stimuli differ only in the initial frequency of the F2 and F3 formant glides. Thus, the distinction between these stimuli lies in spectral cues only.

Figure 1. Schematic spectrogram for a composite /bæ/-/dæ/ stimulus. Hatched bars indicate formant trajectories for /bæ/; dotted bars indicate those for /dæ/.

The authors' mailing address is: Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada T6G 2E7.

When two such signal waveforms with similar fundamental frequencies are added, the result is a waveform that contains the phonetic cues for both /bæ/ and /dæ/. Such a signal is perceptually ambiguous and is perceived as either /bæ/ or /dæ/. Detectability of interference between the two stimulus components and the quality of the resulting percept depends on how well the two component stimuli are aligned in time.

The procedure for aligning the stimuli will be explained with continued reference to the /bæ/-/dæ/ example. To label different segments of the stimulus waveforms, the following notation will be used: /bæ/ designates the entire syllable, and /b/ the extracted formant transitions from that syllable, in our particular case, the first nine glottal periods of /bæ/. The symbol /æ/ indicates the steady-state vowel, in the present example the 10th through last glottal period of the /bæ/ waveform.

The feasibility of aligning the two stimulus components depends on obtaining tokens of the two test stimuli with sufficiently close frequency and time parameters, expected to be common to both stimuli, such as fundamental frequency, formant target frequencies, and durations of corresponding signal segments. For this reason, multiple tokens of /bæ/ and /dæ/ were recorded by a male speaker with approximately equal fundamental frequencies and steady-state vowel formant values. One /bæ/-/dæ/ pair was then selected on the basis of judgments of the similarity of the fundamental frequencies and the compatibility of the formant values of the steady-state vowel segments, as determined from spectrograms of the stimuli and plots of the signal waveforms. These tokens of /bæ/ and /dæ/ were low-pass filtered at 7 kHz to avoid frequency aliasing and then digitized at a sampling rate of 16 kHz using a 10-bit analog-to-digital converter. The digitized waveforms were stored in disk files.

Of the two selected consonantal components to be mixed, the waveform with the lower fundamental frequency, and thus with longer glottal periods, /b/ in this example, was chosen as the standard time frame. To this frame, the /d/ signal was then temporally aligned in the following manner: The digitized /b/ and /d/ waveforms were segmented by manual cursor setting into individual pitch periods at positive-going zero-crossings preceding the major peaks. Thus, separate single-period signals /b_j/ and /d_j/ were obtained, where the index *j* indicates the sequential order of each period, *j* = 1, 2, . . . , 9. Each digitized /d_j/ period was then cross-correlated with its corresponding /b_j/ period over its length of *N_j* sample points. Figure 2 shows plots of the cross-correlation functions *r_j*(*i*),

$$r_j(i) = \frac{1}{N_j - 1} \sum_{n=1}^{N_j} b_{j,n} d_{j,n+i}, \quad (1)$$

for the first six pitch periods. The two signals were considered optimally aligned for the time lag of *i* sampling

Figure 2. Example of covariance traces for the six initial pitch periods for syllables /bæ/ and /dæ/. The arrows indicate the points of maximum correlation between the corresponding pitch periods.

intervals, which yielded a maximum value of *r_j*(*i*). After all of the /d_j/ periods were optimally time-shifted in this fashion, a new /d/ signal was created by concatenating these time-adjusted /d_j/ periods. In this operation, sampling points of zero value were, in effect, inserted between the original /d_j/ periods. The resulting /d/ waveform now matched the original /b/ waveform on a period-for-period basis, as shown in Figure 3.

In exceptional cases, for instance, due to a strong jitter, some of the individual /d_j/ periods may be longer than their /b_j/ counterparts. Then the tail of the /d_j/ segment can be truncated to match the duration of /b_j. For the discussed stimulus pair, such truncation was never required.

The above operations, period extension by adding of sample points of zero value and period shortening by truncation, affect the fundamental frequency of the signal and duration of the signal. The formant frequencies remain unchanged. These operations were used by Connell, Hogan, and Rozsypal (1983) to manipulate the fundamental frequency contours of monosyllabic Chinese words in order to alter their tone patterns, while preserving the original vowel formant values. If the alignment of the tran-

Figure 3. The /b/ and /d/ transition waveforms after temporal alignment of the individual pitch periods of the /d/ waveform.

sition waveform was attempted by compression or expansion of the individual periods, not only the pitch and duration but also the formant frequencies of the modified waveform would be altered.

Next, the /d/ component was amplitude-scaled so that its energy was equal to that of the /b/. This energy was measured as a sum of squares over all of the nine pitch periods of the formant transition /d/ for a total of N points,

$$E_d = \sum_{n=1}^N d_n^2. \quad (2)$$

Signals /b/, /d/, and /æ/ were stored in separate files. These were used as the basic signals for construction of all test stimuli, which were prepared by first adding the /b/ and /d/ formant transitions together point by point in a linear proportion $\alpha b_n + (1-\alpha) d_n$. This resulting composite transition signal had the undesirable property that its energy still depended on the value of the weighting factor α . Since loudness variation can provide an additional cue in discrimination tasks, the amplitudes of all stimuli were adjusted for equal energy. Assuming signals with a mean equal to zero, their energy is equal to their variance, as indicated by Equation 2. Given two sets of corresponding measurements X and Y with respective variances s_x^2 and s_y^2 , the variance of the sum $X + Y$ can be expressed (Ferguson, 1981, Equation 8.4) as

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2rs_x s_y, \quad (3)$$

in which the correlation coefficient r between the two sets X and Y is given (Ferguson, 1981, Equation 8.2) as

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}. \quad (4)$$

From Equation 3 was derived the formula for the energy E_{b+d} of the superimposed /b/ and /d/ waveform,

$$E_{b+d} = \alpha^2 E_b + (1-\alpha)^2 E_d + 2\alpha(1-\alpha)r\sqrt{E_b E_d}, \quad (5)$$

where r is the correlation coefficient between /b/ and /d/, computed as in Equation 4. Because the amplitudes of the added signal waveforms were varied linearly with α , the

energy of the resulting waveform in Equation 5 varied with α quadratically. This relationship would be independent of α only for stimuli of equal energy and perfect correlation, $r = 1$, that is, for identical stimuli. For $r < 1$, this relationship is quadratic, with a minimum at $\alpha = 0.5$. Its value depends on r . Since the /b/ and /d/ formant transitions were initially equated for overall energy, the E_{b+d} curve reaches the same maximum value at the extreme positions $\alpha = 0$ and $\alpha = 1$. A scaling factor equal to the inverse value of E_{b+d} from Equation 5 was used to eliminate differences in energy of the combined transitions stimuli of different α . This energy-equalized signal was then concatenated to the vowel nucleus /æ/. The linear weighting of the /b/ and /d/, along with the energy equalization and the fact that both of the component waveforms converged to the same steady-state vowel, ensured continuity of the envelope at the splicing point between the composite formant transitions and the steady-state vowel. Figure 4 shows spectrograms of the /bæ/-/dæ/ composite stimuli for five different values of α , going from the pure /da/ stimulus on the left to pure /ba/ on the right. Interpretation of these spectrograms can be aided by referring to Figure 1.

APPLICABILITY OF THE TECHNIQUE

Monaural fusion continua have several advantages over synthetic continua. (1) They use real speech, which has all the cues that may be missing in synthesized stimuli. (2) Because natural speech stimuli are used, identification of the endpoint stimuli on the continuum is virtually perfect, guaranteeing a measurable crossover point at the category boundary. And, (3) the monaural fusion paradigm controls a well-defined unidimensional physical variable, that is, the relative intensity of the two stimulus components.

Besides these advantages, the main motivation for developing this technique is to provide an experimental tool for the investigation of the operation of feature detectors or category prototypes. This method makes it possible to stimulate the optimal range of sensitivity of both detectors simultaneously in a controlled manner, since all the stimuli from the continuum contain, in varying intensity ratios, only the natural cues for both categories. These cues are left unmodified, except for some amplitude and

Figure 4. Spectrograms of the composite /bæ/-/dæ/ stimuli for different values of α . From left to right: $\alpha = 0.00$, pure /dæ/; $\alpha = 0.25$; $\alpha = 0.50$; $\alpha = 0.75$; $\alpha = 1.00$, pure /bæ/.

temporal scaling. If it is assumed, for instance, that the /b/ and /d/ cues excite two separate feature detectors, these detectors can be regarded as inhibiting each other if the stimuli are found to fuse, that is, if the subjects hear only one or the other stimulus component. On the other hand, nonfusing stimuli, that is, those that are both heard simultaneously, can be viewed as processed by noninteracting channels.

Figure 5 presents typical individual identification curves of five subjects for the /bæ/-/dæ/ continuum (Stevenson, 1979). All curves show sharp transition from 0% to 100% identification. Postexperimental interviews revealed that, for readily fusing stimuli, the subjects found the identification tests to be easy. They felt confident that they had responded correctly, since each stimulus was generally perceived as a clear instance of one or the other category. This is a major difference from the perception of the interpolated stimuli. With the latter, the boundary stimulus is usually perceived as not being representative of either category, and the subjects are forced to assign it to one or the other category with low confidence.

Repeated listening to a near-boundary stimulus in the fusion experiment can be compared to observing a Necker cube: The subjects can, more or less at will, alternate between the two percepts. A slight increase in noisiness, perceived by a few subjects in the stimuli from the category boundary region, is probably due either to imperfection of period alignment or to "beats" between close frequency components of the two sets of formant transitions constituting the stimulus. But even these stimuli are classified by those subjects with confidence.

On the other hand, in the constant-step discrimination experiments, $\Delta\alpha = 0.05$, most subjects experience a lot of frustration due to their lack of confidence in giving "different" responses. Stimuli within the category are discriminated at the chance level. A better method to test discrimination with the composite stimuli appears to be

the fixed-standard discrimination paradigm, in which, at least for some of the stimulus pairs, subjects can detect the difference with confidence.

Although the technique so far has been discussed solely in terms of the /bæ/-/dæ/ pair, the procedure, as well as its perceptual effects, is quite general and can be applied to various other speech waveforms. For instance, Stevenson (1979) employed the same technique, in experiments of lesser or greater extent, also for stimulus pairs /bi/-/di/, /bet/-/det/, /ra/-/la/, /bæ/-/ræ/, and /i/-/æ/, and even for a triplet /ba/-/da/-/ga/.

The most suitable stimulus types for fusion are consonant-vowel (CV) and consonant-vowel-consonant (CVC) monosyllables, in which the segments differ in spectral cues only, such as pairs /bæ/-/dæ/, /bi/-/di/, and /bet/-/det/. In listening to the /bæ/-/ræ/ combination, subjects report hearing "bræ" at the transition from /bæ/ to /ræ/. In the /ra/-/la/ experiment, subjects perceive both initial consonants simultaneously; some even reported hearing "bra" or "bla." It appears that vowels do not fuse as readily as consonants: For the /i/-/æ/ combination, both vowels could always be heard simultaneously, even when the level of the weaker vowel was 25 dB below the stronger one. Thus, the degree of fusion may become one of possible measures of the categoricity of the continuum.

The potential applications of this technique to other speech sounds can be fairly extensive. With reference to the International Phonetic Association (1949) consonant chart, the stimulus types selected for fusion can be chosen from the 11 designated place-of-articulation categories, provided that they are in the same manner class. Tests can be carried out also for stimulus types from certain different manner classes, given a fixed place of articulation. For example, tests consisting of plosives with either ejectives or implosives can be constructed. Similarly, a comparison of voiced plosives with nasals may be attempted, even though the duration of the plosive burst and the nasal murmur are not equal. Other central and lateral approximants besides /r/ and /l/ can also be used.

The method is equally well applicable to synthetic speech stimuli (Repp, 1981). In this case, the procedure of temporal alignment is even simpler, because the pitch contours of the two components, exactly controlled in the synthesis, can be made identical.

Certain limitations on the type of speech signals that can be combined to produce monaural fusion are apparent: Only stimuli of similar temporal structure can be employed in this method. Combinations of stimuli differing in voice-onset time and combinations of a plosive with a fricative have not yet been tested, and most likely may not fuse.

As a possible extension of the described method, component combination in more than one position within the stimulus deserves mention. For instance, in a CVC stimulus, both the initial and final consonant can be mixed independently from two components, yielding four alter-

Figure 5. Identification curves of the composite /bæ/-/dæ/ stimuli as a function of α for five subjects.

native responses. Another possibility is to scale the intensity of only a part, rather than the full length, of a particular segment, which yields information about the relevance of the subsegments in question. Composite stimuli can be applied also in testing binaural interaction: A stimulus pair with intensity ratio α can be presented to one ear, and the complementary pair with ratio $(1 - \alpha)$ to the other (Stevenson, 1979).

REFERENCES

- CONNELL, B. A., HOGAN, J. T., & ROZSYPAL, A. J. (1983). Experimental evidence of interaction between tone and intonation in Mandarin Chinese. *Journal of Phonetics*, **11**, 337-351.
- FERGUSON, G. A. (1981). *Statistical analysis in psychology and education* (5th ed.). New York: McGraw-Hill.
- INTERNATIONAL PHONETIC ASSOCIATION. (1949). *The principles of the International Phonetic Association*. London: Author.
- REPP, B. H. (1981). Perceptual equivalence of two kinds of ambiguous speech stimuli. *Bulletin of the Psychonomic Society*, **18**, 12-14.
- ROZSYPAL, A. J. (1976). Digital gating of speech signals. *Language and Speech*, **19**, 57-74.
- STEVENSON, D. C. (1979). *Categorical perception and selective adaptation in speech*. Unpublished doctoral dissertation, University of Alberta, Alberta, Canada.
- STEVENSON, D. C., & STEPHENS, R. C. (1978). A programming system for psychoacoustic experimentation. *Proceedings of the 11th DE-CUS Symposium*, Ottawa, Canada, 837-844.

(Manuscript received October 13, 1984;
revision accepted for publication February 12, 1985.)