General description

The boot-beet and said-shed continua were created in 2001. Each continuum has 13 stimuli and there are a total of four continua (one each of beet-boot and said-shed for a female and a male speaker). There are two kinds of files in this archive, parameter and audio. The parameter files have the .doc extension, but they are text files that serve as input to a Klatt synthesizer rather than Microsoft Word files. The audio files are Microsoft .wav files.

The naming convention for the parameter files is

<gender code>_<s_sh,u_i>_<stimulus number>.doc

where "gender code" is "m" for the male continua and "f" for the female continua, and "stimulus number" is 01-13.

The naming convention for the audio files is

<gender code>_<said_shed, boot_beet>_<stimulus number>_ms.wav

where "gender code" is "m" for the male continua and "f" for the female continua, and "stimulus number" is 01-13. The "ms" in the file names indicates that the file format is Microsoft .wav rather than Klatt .wav, which is commonly used in the Speech Communication Group.

Synthesis procedure

Because it was desired to have stimuli that sounded as much like natural human speech as possible, copy synthesis was used to create beet/boot and said/shed continua. Details of the copy synthesis method can be found in Hanson (1995). Recordings were made of one male and one female subject saying "a beet," "a boot," "a said", and "a shed". The subjects were asked to read the phrases as naturally as possible. They read the phrases from a sheet of paper. Each phrase was read three times, in a semi-random order. These recordings were made to DAT in a sound-attenuated chamber at a sampling rate of 48,000 samples/s. After being transferred to a computer, the recordings of "a beet" and "a boot" were downsampled to 11,000 samples/s, and the recordings of "a said" and "a shed" were downsampled to 16,000 samples/s. One token of each of the four phrases (per speaker) was chosen as the basis of the copy synthesis. These choices were based on factors such as formant transitions, durations, amplitudes, and regularity of voicing. Because the methods for synthesizing vowels and sibilants are somewhat different, we describe the synthesis of each continuum separately.

"beet"-"boot" continuum. In the first stage of the synthesis procedure, vowels i_S and u_S were synthesized to match the vowel portions of the natural utterances "beet" and "boot." A Klatt formant synthesizer was used (Klatt and Klatt, 1990). Parameters that varied with

time included formants F1-F4, their associated bandwidths, F0, and the voice-source parameters AV (amplitude of voicing), AH (amplitude of aspiration), and TL (spectral tilt). These parameters were adjusted until the experimenter was convinced that the synthesized vowels sounded identical to the natural vowels, and that spectra and spectrograms of the synthesized vowels closely matched those of the natural. Note that this procedure matches the formant-frequency variation and transitions that naturally occur over the course of a syllable.

In the second stage of the procedure, the synthesizer parameters for i_S and u_S were adjusted to create endpoint stimuli i_E and u_E that differed only in their F2 and F3 trajectories. This adjustment involved changing the F0, F1, F4-F6, B1-B6, AV, AH, and TL trajectories to have the same values and durations for both i_E and u_E. Note that, in general, these adjustments were minor, as should be expected. For example, because both /i/ and /u/ are high vowels, there is little difference in the F1 trajectory. The F2 and F3 trajectories were adjusted to have the same duration as the other parameters, but otherwise retained their original values. Informal listening tests verified that the stimuli i_E and u_E sounded natural and as though they were spoken by the original speakers. Again, we note that the formant frequencies vary in a natural way over the course of the vowels.

In the third stage of the synthesis, a continuum was created between i_E and u_E by linearly interpolating the F2 and F3 trajectories. Seven intermediate stimuli were created. The step-size between stimuli varies depending on time, and the formant transitions that occur at vowel onset and offset are intermediate between those of i_E and u_E. In addition to the seven intermediate stimuli, two additional stimuli were added at each end of the continuum to extend beyond the parameter values of i_E and u_E. The step sizes used for these stimuli were the same as those used for the intermediate stimuli. Figure 1 illustrates the formant-trajectory continua for F2 and F3.

In the final stage of the synthesis, an initial /b/ (from closure to voice onset) and a final /t/ (from closure to shortly past the release) were excised from one of the natural tokens of "beet" and "boot", and concatenated with the 13 stimuli of the vowel continuum to create the beet-boot continuum. This procedure was relatively simple because the vowels were in a stop environment, and thus there were no problems with discontinuities in the waveforms. Our reasoning for doing it this way, rather than also synthesizing the /b/ and /t/, was that it simplified the procedure, and, as long as the resulting stimuli sound natural, it is irrelevant if the initial and final stops are synthesized or natural. Informal listening tests among members of our laboratory verified that the stimuli do indeed sound natural.

"said"-"shed" continuum. The usual approach for synthesizing a /s/-/ʃ/ continuum is to vary formant frequencies, e.g., to shift F3 from a value typical for /ʃ/ to a higher value that is typical of the lowest frication-excited resonance of /s/. Based on acoustic models of speech production (e.g., Stevens, 1998), we observe that this method does not truly model the acoustic contrast between /s/ and /ʃ/, which is less due to a contrast in formant frequencies than to a contrast in formant amplitudes. The contrast in formant amplitudes results from a difference in which formants are affiliated with the front cavity (and thus are strongly excited by frication) and which with the back cavity (weakly excited).

Acoustic theory predicts that the lowest front-cavity resonance for /ʃ/ will be F3, while the lowest front-cavity resonance for /s/ will be F4 or F5, and observations of natural data, including those recorded for this study, bear that out (Keyser and Stevens, in press). Thus, the method we used to synthesize our s-ʃ continuum was to vary formant amplitudes rather than formant frequencies.

Sibilants s_S and $ʃ_S$ were synthesized to match the sibilant portions of the natural utterances "said" and "shed". The Klatt formant synthesizer was again used, and because the sound source was the frication source, it was required that the parallel mode of the synthesizer be used. In the parallel mode of the synthesizer, formant amplitude is controlled by parameters A2F-A6F which set the amplitudes of the formants. (In the cascade mode, typically used for vowels, formant amplitude is largely controlled by formant frequency and bandwidth.) The synthesized sibilants s_S and $ʃ_S$ were matched to the natural sibilants by adjusting the formant frequency and amplitude values until the experimenter felt that the power spectrum of the synthesized sibiliant was a good match to that of the natural sibilant. Note that the synthesizer parameters did not vary with time, because little formant variation was observed in the natural sibilants.

In the second stage of the procedure, the synthesizer parameters for s_S and $ʃ_S$ were adjusted to create endpoint stimuli s_E and $ʃ_E$ that differed only in A2F-A6F. This adjustment involved changing F2-F6 to have the same values for both sibilants. Only minor adjustments were necessary, and informal listening tests confirmed that the stimuli s_S and $ʃ_S$ sounded natural and of good quality.

In the third stage of the synthesis, a continuum was created between s_E and $ʃ_E$ by interpolating the amplitudes of F2-F6, i.e., the parameters A2F-A6F. Seven intermediate stimuli were created. The stepsize between stimuli was not linear, for two reasons. First, the synthesizer required integer values for the parameters, and therefore a strictly linear interpolation was not possible. Second, pilot tests showed that normal-hearing subjects perceived more than half the stimuli as /s/; therefore, adjustments were made in stepsize to result in normal-hearing subjects hearing half the stimuli as /s/ and half as /ʃ/. In addition to the seven intermediate stimuli, two additional stimuli were added at each end of the continuum to extend beyond the parameter values of s_E and $ʃ_E$. For some of the formants, the step sizes used for these stimuli were significantly larger than those used for the intermediate stimuli, in order for them to sound less than optimal to normal-hearing subjects. The formant frequency and amplitude values that were used are summarized in Table 1.

In the final stage of the synthesis, a final /ɛd/ (from voice onset of the /ɛ/ to shortly past the release of the /d/) were excised from one of the natural tokens of "said" and "shed", and concatenated with the 13 stimuli of the sibilant continuum to create the said-shed continuum. This procedure was easily accomplished because (1) there is not much difference in formant transitions following /s/ and /ʃ/, and what there are are not strong cues to their contrast, and (2) there is usually a transition region with very low amplitude be-

tween unvoiced sibilants and a following vowel, so there were no issues of waveform discontinuities. In addition, the use of the same formant frequencies for both /s/ and /ʃ/ simplified the match up between the synthesized sibilants and the natural vowel. Informal listening tests among members of our laboratory verified that the stimuli do indeed sound natural.

References

Hanson, H. M. (1995). "Synthesis of female speech using the Klatt synthesizer," *Speech Comm. Group Working Papers, X*, Res. Lab. of Elec., MIT.

Keyser, S. J. and K. N. Stevens (in press). "Enhancement and overlap in the speech chain," *Language*.

Klatt, D. H. and L. C. Klatt (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acous. Soc. Am*., 87, pp. 820-857.

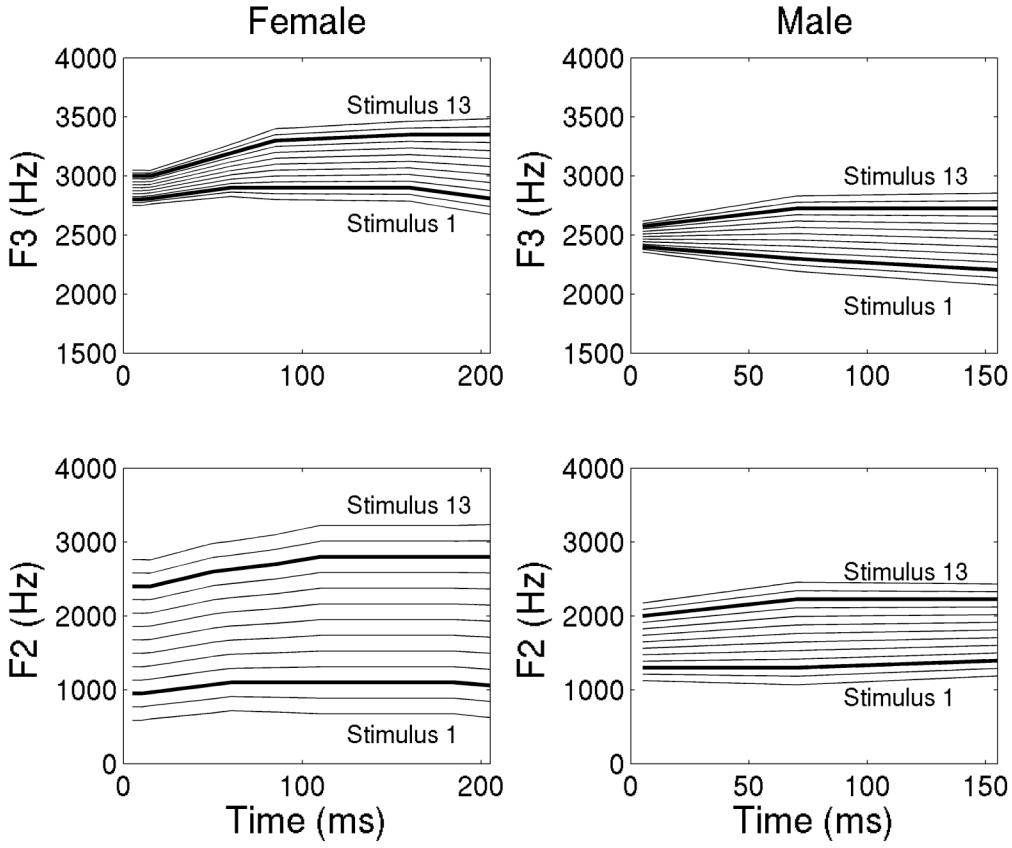

Fig. 1. Trajectories for F2 and F3 of the /i/-/u/ continua. The heavier lines in the figures indicate the trajectories for the endpoint stimuli.

Table 1. Formant amplitude and frequency values used in the synthesis of the sibilant continua. Formant amplitudes are given in dB and formant frequencies are given in Hz. Stimuli 3 and 11, indicated with bold typeface, are the endpoint stimuli.

	Female					Male				
Stimulus	A2F (F2 = 1900)	A3F (F3 = 3200)	A4F (F4 = 4700)	A5F (F5 = 6000)	A6F (F6 = 7200)	A2F (F2 = 2500)	A3F (F3 = 3750)	A4F (F4 = 5000)	A5F (F5 = 5800)	A6F (F6 = 6800)
1	33	34	43	80	80	17	29	79	79	74
2	34	37	44	70	70	27	34	72	72	67
3	**35**	**40**	**45**	**60**	**60**	**35**	**39**	**67**	**67**	**62**
4	36	42	46	69	58	39	43	63	63	58
5	37	45	46	57	57	44	47	60	60	55
6	37	47	47	56	56	48	51	58	58	53
7	38	50	47	55	55	53	55	56	56	51
8	38	52	48	54	54	59	59	54	54	49
9	39	55	48	53	53	62	62	52	52	47
10	39	62	49	51	51	64	64	49	49	44
11	**40**	**65**	**50**	**50**	**50**	**66**	**66**	**47**	**47**	**42**
12	41	70	51	48	48	68	68	46	46	41
13	42	75	52	46	46	70	70	44	44	39