English listeners' perception of Polish alveopalatal and retroflex voiceless sibilants: A pilot study[1]

Grant McGuire[2]
University of California, Berkeley

Abstact: This paper describes results from a series of brief pilot experiments exploring the perception of Polish alveopalatal and retroflex voiceless sibilants by native speakers of English. The goal of these experiments was to examine the suitability of a two-dimensional stimulus set for use in a series of training experiments. The stimulus set, consisting of CV syllables varying from alveopalatal to retroflex in two dimensions, fricative noise and vocalic cues, was created by modifying naturally produced tokens. Generally, English listeners were sensitive to distinctions in the stimuli on limited basis. Specifically, and unlike a native speaker of Polish, English listeners rely solely on vocalic dimension to categorize the stimuli and ignore fricative variation. However, with brief training, attention to that dimension was possible, with minor improvement.

1.0 Introduction

Training studies using adult listeners allow for an exploration of category formation and relevant theoretical issues. One important factor in perceptual learning is an understanding of which perceptual cues aid in category identification and how they are weighed by listeners. This paper examines the suitability of a stimulus design method and a phonetic contrast, Polish post-alveolar voiceless sibilants, for use in a series of perceptual learning experiments examining cue learning.

1.1 Cue learning and weighting in speech perception

Adults can easily recognize perceptual categories in their native language under most conditions. In order to perform this feat, listeners must be able to recognize and weigh relevant cues in the speech signal. Thus an important factor in category acquisition is learning such cues and their relative weights in the proper contexts (Nittrouer and Miller 1997, Nittrouer 2002, Hazan and Barret 2000, Mayo and Turk 2005).

1 This work was supported by NIDCD grant R01 DC004421. This data was first reported as part of a doctoral dissertation by the author. Thanks go to Keith Johnson, Mary Beckman, Susan Nittrouer, and menbers of the Berkeley phonetics and phonology reading group for valuable input.

2 Contact: grantmcguire@berkeley.edu

Several studies have examined differences in cue use by children and adults. For example, Nittrouer and colleagues have done considerable work examining children's perception of place cues in fricatives (Nittrouer and Studdert Kennedy 1987, Nittrouer 1992, Nittrouer and Miller 1995, Nittrouer 2002). Generally, this work uses synthetic or natural stimuli arranged into continua where either transition or fricative noise cues are minimized or are varied independently. Nittrouer and Miller (1997) and Nittrouer (2002) found that young English learning children (3.5 years old) generally weigh transition cues heavily when discriminating place in sibilant fricatives (/s/ and /ʃ/), gradually reaching English speaking adult-like weighting (fricative noise more informative than transition) around 7-8 years old. Nittrouer (2002) demonstrated that for the /f/ ~ /θ/ distinction English learning children predictably used cues like English speaking adults, who weigh formant transition much more heavily than fricative noise. Importantly, this work shows that as children age they must differentially weigh cues in relation to what is the most reliable cue for a given contrast; this process apparently takes many years (≈ 4) and amounts to a gradual refinement of the perceptual system.

In a similar vein, Hazan and Barrett (2000) examined English learning children 6-12 years old and their ability to perceive several phonemic contrasts, with and without all available cues. The contrasts ranged from robust to fragile as based on frequency in the world's languages, respectively /k/~/ɡ/, /d/~/ɡ/, /s/~/z/, /s/~/ʃ/, and were represented by minimal pairs. All contrasts (except /s/ ~ /z/) were tested in such a way that static and dynamic cues were in isolation as well as in their natural combinations. The results show a general increase in categorization accuracy, but 12 year olds were still not quite at adult-like levels. Generally, the children showed more inconsistency in isolated cue conditions compared to combined cue conditions; adults, however, seem to be easily able to switch to different perceptual strategies and cue-use demands.

Taken together, the results from these various studies show that younger children are not as sophisticated in their use of redundant information as adults and that their category boundaries are somewhat fuzzy. Adults seem to be able to use a wide variety of cues and perceptual strategies, both in conjunction and independently, while children are much more limited.

However, despite this research, the role of cue acquisition and weighting in adult learners has not been addressed to the same extent. Recently, though, several authors have begun to explore this issue. First, Francis, Baldwin, and Nusbaum (2000) trained English listeners to selectively attend independently to stop burst and formant transition information of native stops. They found an increase in sensitivity to the trained cues with a concurrent loss of sensitivity to the untrained cue. A second study by Francis and Nusbaum (2002) trained English listeners to distinguish Korean phonation type contrasts. A multi-dimensional scaling analysis demonstrated that subjects could learn to use novel cues when necessary to make a contrast and weigh important cues as necessary. A third study, Guion and Pederson (2007), explored Mandarin native speakers' and learners' use of tone. The authors demonstrated that only with considerable experience do later learners of Mandarin use tonal cues in the same way that native listeners do while new (early) learners do not show the same patterns as native listeners.

Overall, these studies demonstrate that in order to learn a contrast and to behave as a native listener, the proper cues must be attended to and weighted appropriately. For native cues, the learning process takes many years, with children near puberty still showing not-quite adult-like perceptual abilities. Similarly, experimental evidence with adults shows that with proper training subjects can be directed to attend to relevant dimensions and can also learn novel cues if necessary for categorizing.

1.2 Polish sibilants

A phonetic contrast must be sufficiently difficult to for subjects to show improvement over the course of training. As suggested by Best et al. (2001) listeners confronted with a new contrast will map them onto their own native categories if possible. When a contrast is subsumed under the variation of a native contrast, then that contrast will be very difficult for listeners to distinguish. For English listeners, one such non-native contrast is the Polish alveopalatal /ɕ/ and retroflex /ʂ/ sibilant categories[3], which they

3 There is some disagreement on the exact phonetic transcription of these sounds. See Nowak (2006) and Zygis and Hamman (2003) for discussions. I follow the conventions found in Nowak (2006).

generally collapse under the native category /ʃ/ (Lisker 2001). This distinction is also found in many varieties of Mandarin Chinese including the Beijing dialect and the PRC Putonghua standard that is based on it.

Although both sounds are similar to English /ʃ/, they are both articulatorily distinct from it. Ladefoged and Maddieson (1996) report that while both English /ʃ/ and Polish /ʂ/ have similar constriction locations and widths along with concurrent lip-rounding, /ʂ/ has a much flatter tongue shape. For /ʃ/ and /ɕ/, they note that the tongue blade and body are higher for /ɕ/ and that both exhibit lip rounding and have very similar place of articulation to the retroflex. In comparing the Polish sibilants with the Mandarin sibilants, they find very similar articulation strategies for both sibilants with the exception that Mandarin /ʂ/ does not exhibit lip rounding but has a larger sub-lingual cavity than its Polish equivalent.

For native speakers of Polish this distinction has several cues with the primary ones being fricative pole frequency and F2 onset (Lisker 2001, Zygis and Hamann 2003, Nowak 2006), along with slight post-consonantal vowel quality differences (Nowak, 2006). Specifically, Nowak (2006) demonstrated that proper formant transition information is necessary for native speakers identifying syllables containing these sibilants. However, isolated fricatives could also be reliably identified. He suggests that subjects used very different perceptual strategies in the different conditions, possibly perceiving the isolated fricatives as non-speech. In any event, it is clear that both fricative noise and formant transition information are used by Polish listeners for identification of these fricatives.

In contrast, English listeners show very poor discrimination of this contrast (Lisker 2001). Lisker's study, using brief training, found that English speakers could not discriminate these sounds above chance in the context of full syllables, but could discriminate if presented with either the fricatives alone or the following vowels in isolation. Lisker suggests a non-speech mode of perception to account for the discrep-

ancies between full-syllable and isolated segment conditions. However, the fact that English listeners could, with very little training, discriminate the isolated segments suggests that more extensive training could expand on these abilities.

Given these findings, this contrast seems to be an ideal candidate for studying perceptual learning. In order to compare the two sources of identification information (fricative noise and vocalic transition) it is necessary to have a stimulus set that varies in both dimensions and has sufficient internal structure such that changes within as well as across categories can be examined. The following stimulus design and experiments address these concerns.

2.0 Experiments

2.1 Stimuli

The stimuli for the following experiments consisted of Polish alveopalatal and retroflex sibilant consonants followed by [a] in a two-dimensional space varying by fricative noise in one dimension and by vocalic transition information in the other dimension. This space was produced by interpolating modified, naturally produced examples of the desired syllables[4]. This method of construction was chosen as it preserves the full richness of the auditory signal.

Specifically, several productions of [ʂa] and [ɕa] were recorded in a sound-proof booth by a male native speaker of Polish using a head-mounted microphone and a Marantz PMD670 solid state recorder at 44.1 kHz sampling rate. One example of each syllable was selected based on clarity and similarity to the acoustic analyses of Polish fricatives reported in Nowak (2006). The selected retroflex syllable had a peak located at 2890 Hz and an F2 onset at 1420 Hz with a midpoint F2 of 1280 Hz. The alveopalatal had a peak located at 3890 Hz with an F2 onset of 1720 Hz and a midpoint F2 of 1320 Hz.

4 See appendix for a discussion and example of the scripts used.

Each syllable was split in two at the boundary of the fricative and vocalic portions, determined by the onset of voicing. The two fricatives were brought to the same length by excising 32ms from [ɕ] in four 8ms chunks located at 20% intervals of the total length. The vowels were modified using Praat (Boersma and Weenink 2002) to have the same length, pitch, and RMS through PSOLA resynthesis. Both the fricative and vowel portions were then separately interpolated to form fricative and vowel continua consisting of 10 steps where each step was one of ten graded proportions in terms of intensity. That is, fricative step 0 consisted of 9/9 [ɕ] and 0/9 [ʂ], while step 1 consisted of 8/9 [ɕ] and 1/9 [ʂ], step 2 was 7/9 [ɕ] and 2/9 [ʂ], etc.. Figure 1 displays spectra of selected fricative steps (20ms hamming window from the center of the fricative, cepstral-smoothed 500Hz bandwith) and Table 1 displays vowel formant measures taken at 25ms and 100ms from onset of voicing.

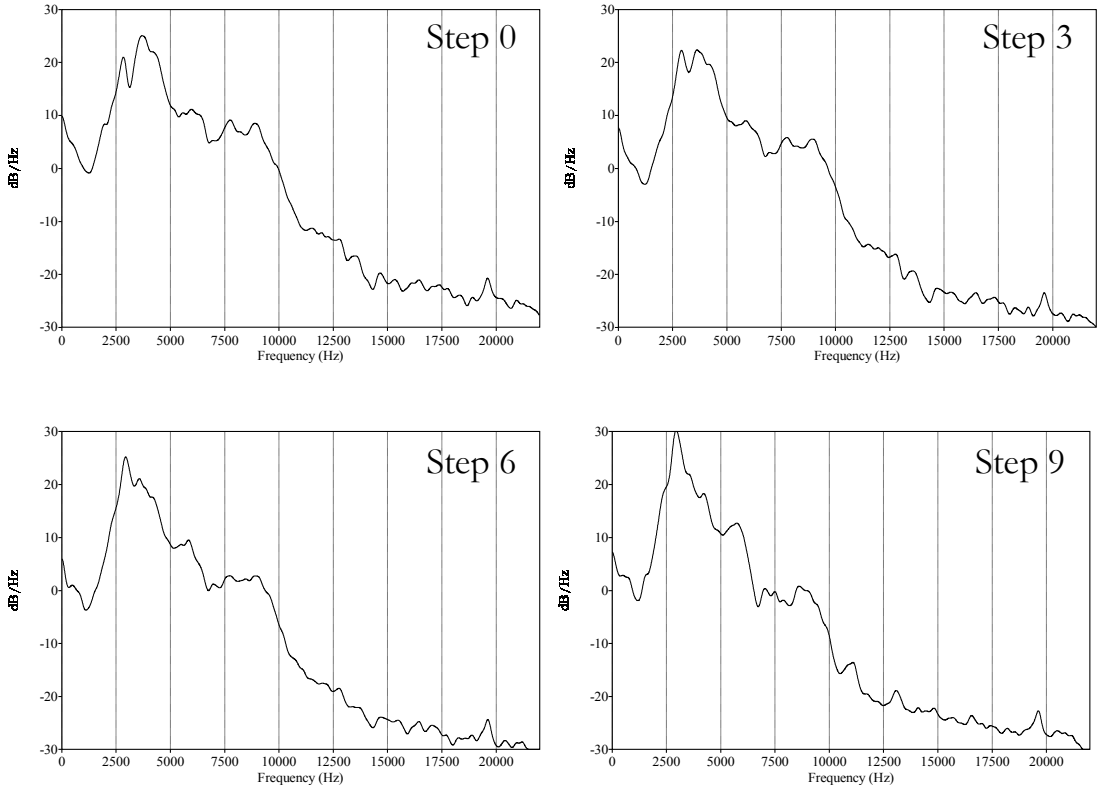

Figure 1: Spectra of fricative step 0 (fully alveopalatal), step 3, step 6, and step 9 (fully retroflex).

Step	25ms		100ms	
	F1	**F2**	**F1**	**F2**
0	680	1629	775	1315
1	684	1624	778	1317
2	690	1617	777	1315
3	701	1607	777	1311
4	716	1597	776	1308
5	734	1581	775	1304
6	752	1556	776	1300
7	763	1513	775	1295
8	765	1455	776	1289
9	762	1392	777	1283

Table 1: Formant values (in Hz) for each step of the vowel continuum as measured at +25ms and +100ms from the onset of voicing.

Although interpolated vowel signals usually results in the percept of multiple vowels rather than a unified percept, equalizing the duration and pitch contour avoids this complication (Scheffers 1982, Zwicker 1984, Bregman 1994). Moreover, the closely spaced formants from each separate sound should be perceived as a single formant based on the center of gravity effect described by Chistovich and Lublinskaya (1979).

CV syllables were produced by concatenating each fricative with each vowel yielding 100 tokens varying in two dimensions, vowel transition and fricative noise. The concatenated "natural" endpoint syllables are shown in Figure 2. The 10 X 10 stimulus set is graphically represented in Figure 3.

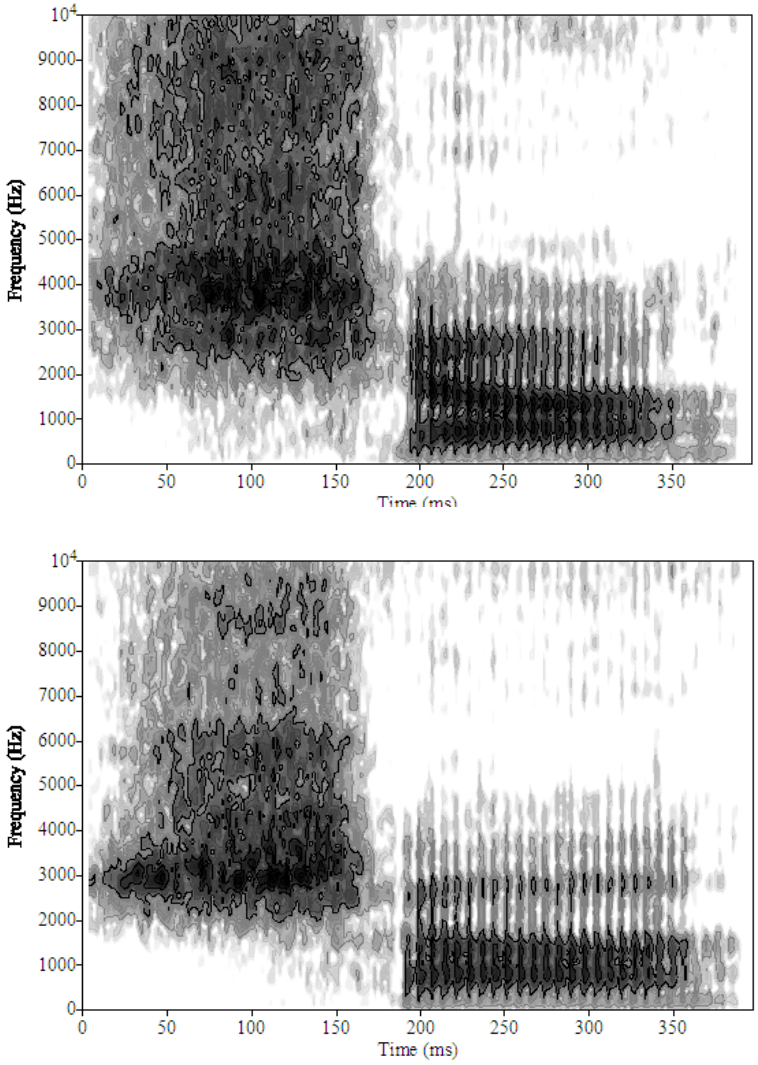

Figure 2: Spectrograms of the fully alveopalatal (top) and fully retroflex (bottom) syllables.

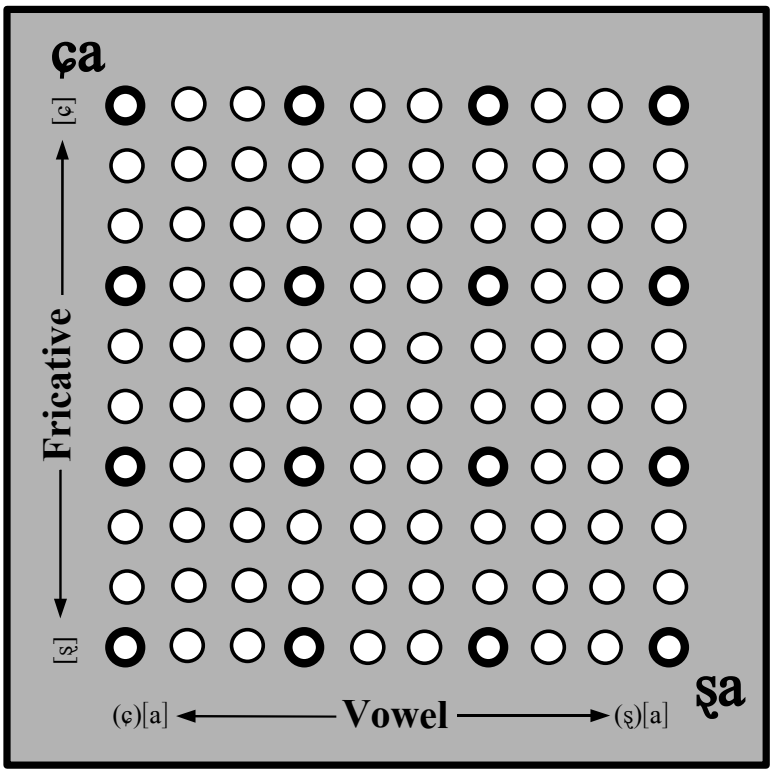

Figure 3: Stimuli design. Each circle represents a particular combination consonant and vowel from each continuum. Circles with darker outlines represent a subset of tokens for orthographic labeling (see text.)

2.2 A Polish listener's perception

In order to assure that the stimuli still represent consistent linguistic categories after modification, a native speaker of Polish was asked to label them. The participant was the same speaker who produced the stimuli and is a trained linguistic phonetician who has studied Polish fricatives, but was not aware of how the stimuli had been manipulated. Each stimulus was presented in random order, in a single block (n=100).

Three blocks were presented for a total of 300 trials. The subject was asked to label each stimulus by responding using a five-button box with the leftmost button labeled *sza* (retroflex) and the rightmost labeled *śa* (alveopalatal). The interval between trials was 3s, there was no feedback given.

The labeling results show a clear categorical distinction between the two categories (see Figure 4). The boundary along the fricative dimension is approximately between fricative step 4 (f4) and fricative step 5 (f5), the center of that dimension. The vocalic boundary is shifted considerably towards the retroflex end of that dimension, around vocalic step 6 (v6) and vocalic step 7 (v7).

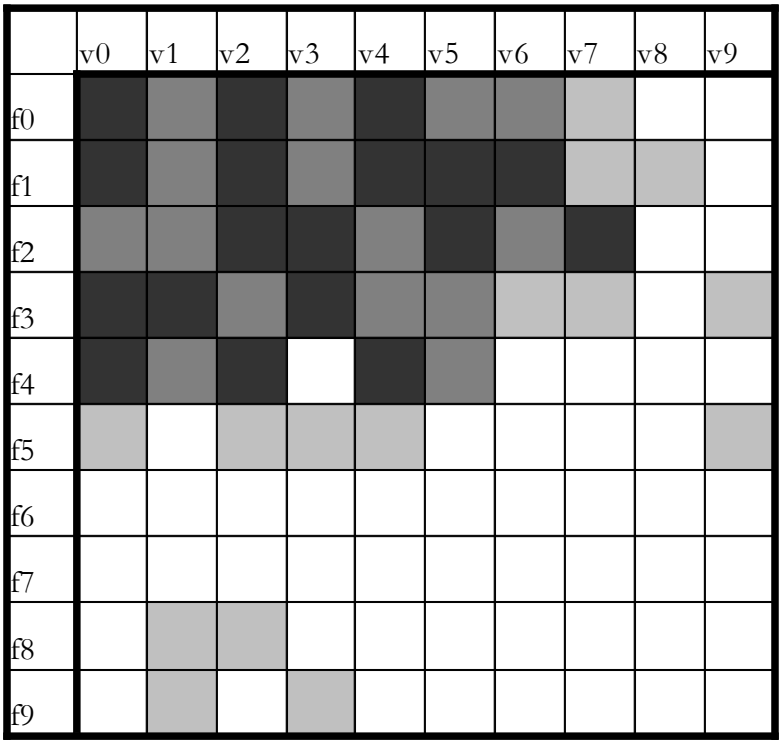

Figure 4: Labeling results of the stimuli set by a native speaker. Black indicates for 3/3 trials the stimulus was labeled as alveopalatal; dark gray, 2/3; light gray 1/3; white, 0/3, i.e. labeled 100% as retroflex.

These labeling results indicate a clear categorical boundary for this listener. Both dimensions were used for classification, though the boundary was not symmetrical as many more tokens were labeled as retroflex than alveopalatal. It is difficult to draw too many conclusions beyond this from one listener, especially since the listener was a trained linguist. Important to the questions at hand, however, is how English listeners perceive these sounds. The following studies attempt to shed light on that question.

2.3 English listeners' perception of the Polish stimuli

In order to establish the suitability of these stimuli for a training experiment, English listeners' perception of the stimuli was examined with the following questions in mind: 1) Do English listeners uniformly assimilate the Polish contrast to English /ʃ/, or are they sensitive to differences and able to categorically label differences? 2) If English listeners are not sensitive to the differences in the stimuli, what amount of training would be necessary to achieve categorical perception? In order to answer these questions, two studies were run. In the first, listeners labeled the stimuli using English orthography and in the second subjects were briefly trained to categorize the stimuli.

2.3.1 English orthographic labeling

To ascertain English speakers' judgments concerning the stimuli, five native speakers of American English labeled a subset of the stimuli using English orthography. The subset consisted of 16 equally spaced stimuli from the larger set of 100. The stimuli in this subset sample the full range of the fricative and vowel transition dimensions separating Polish [ʂ] and [ɕ] (see fig. 1). Listeners labeled each token five times. Tokens were presented in random order (five repetitions of the list of 16 tokens or one randomization of the 16*5 trials) and the listeners entered their responses on a computer keyboard. Each response was presented presented back to the subject on the subject's computer screen. After two seconds a new sound was presented along with a blank screen.

Table 2 shows the labels used by each subject and their frequency of use. The label *sha* was the most commonly used label for all subjects. The second most common labels were *shya* and *shia,* and only subject 101 did not use one of these two labels. Additionally, subject 100 used a large number of different labels (9), although only a handful of these were used frequently. The other subjects used fewer labels, with 104 only using 3.

Label	s100	s101	s102	s103	s104	Total
sha	*44*	*47*	*39*	*40*	*53*	**223**
shya				*26*	*25*	**51**
shia	*20*		*25*	*1*		**46**
shot		*20*	*1*			**21**
schia	*6*			*10*		**16**
ssha		*10*				**10**
sa			*8*			**8**
scha	*5*			*2*		**7**
tza			*5*			**5**
sha3					*2*	**2**
shaw			*1*	*1*		**2**
shz			*1*			**1**
tsha		*1*				**1**
zha	*1*					**1**
shja	*1*					**1**
shout		*1*				**1**
chia	*1*					**1**
s	*1*					**1**
sch	*1*					**1**
sja		*1*				**1**

Table 2: Labeling responses from the five subjects.

All subjects used *sha* as a label for all stimuli, except for subject 104, who never labeled stimulus f0-v0 (alveopalatal fricative + alveopalatal vowel) as *sha*, but otherwise used it extensively. Four of the five subjects indicated a distinction along the vowel continuum using the label *shya* or *shia*, typically used for stimuli with vowels v0 and v3 (i.e. the alveopalatal end of the continuum). Other labels were used extensively, but none showed a clear pattern of use. The lone exception to this is a single subject (101) who used *ssha* as a label for f0 and f3 tokens and did not use any label to indicate a distinction along the vocalic dimension. This subject also used the label *shot* frequently (20 times), though there was no discernible pattern Further exceptional behavior from this subject will be discussed in the next section. Table 2.3 shows the use of *sha* and *shia, shya,* or *ssha* by all subjects.

Subject 100									
"sha"	v0	v3	v6	v9	"shia"	v0	v3	v6	v9
f0	*2*	*1*	*3*	*5*	f0	*2*	*2*	*1*	*0*
f3	*1*	*2*	*4*	*4*	f3	*3*	*0*	*1*	*1*
f6	*2*	*1*	*4*	*5*	f6	*1*	*2*	*0*	*0*
f9	*1*	*1*	*2*	*6*	f9	*3*	*2*	*2*	*0*

Subject 101									
"sha"	v0	v3	v6	v9	"ssha"	v0	v3	v6	v9
f0	*2*	*2*	*1*	*2*	f0	*2*	*0*	*1*	*1*
f3	*2*	*2*	*3*	*4*	f3	*3*	*2*	*1*	*0*
f6	*3*	*4*	*4*	*4*	f6	*0*	*0*	*0*	*0*
f9	*4*	*4*	*3*	*3*	f9	*0*	*0*	*0*	*0*

Subject 102									
"sha"	v0	v3	v6	v9	"shia"	v0	v3	v6	v9
f0	*2*	*2*	*2*	*4*	f0	*3*	*3*	*2*	*0*
f3	*2*	*1*	*3*	*2*	f3	*2*	*3*	*1*	*0*
f6	*3*	*1*	*4*	*2*	f6	*2*	*3*	*0*	*0*
f9	*2*	*1*	*4*	*4*	f9	*3*	*3*	*0*	*0*

Subject 103									
"sha"	v0	v3	v6	v9	"shya"	v0	v3	v6	v9
f0	*2*	*2*	*2*	*4*	f0	*3*	*3*	*0*	*0*
f3	*2*	*1*	*3*	*3*	f3	*4*	*2*	*0*	*1*
f6	*1*	*2*	*4*	*3*	f6	*3*	*4*	*1*	*0*
f9	*2*	*2*	*3*	*4*	f9	*2*	*3*	*0*	*0*

Subject 104									
"sha"	v0	v3	v6	v9	"shya"	v0	v3	v6	v9
f0	*0*	*2*	*4*	*5*	f0	*5*	*3*	*0*	*0*
f3	*2*	*4*	*5*	*5*	f3	*3*	*1*	*0*	*0*
f6	*1*	*3*	*5*	*5*	f6	*4*	*2*	*0*	*0*
f9	*1*	*2*	*4*	*5*	f9	*4*	*3*	*0*	*0*

Table 3: Response tallies for each subject for the label "sha" and any other label consistently used.

Overall, these subjects are using labels indicating the English palatal glide for stimuli having the alveopalatal formant transitions, although all stimuli may be labeled as *sha*. However, subjects, with one exception, are not making a distinction along the fricative dimension. This would seem to indicate that most English listeners are able to perceive a categorical difference in only the vocalic dimension. Nevertheless, it is also quite possible that limitations in English orthographic representations make it possible to represent the alveopalatal transitions, but not the fricative noise.

Note that the one subject who did label the alveopalatal fricatives differently resorted to a novel representation, *ssha*. This subject's other unique label, *shot,* was not used consistently. Interestingly, this subject has significant experience with Turkish, speaking with native proficiency. The extent to which this accounts for the unique labeling pattern is unknown[5].

In order to discern whether the variation in fricative noise can be used by English listeners further exploration is necessary. Consequently. the following experiment is designed to train listeners to associate abstract labels with specific regions of the stimulus set.

2.3.2 English labeling with brief training

In this experiment subjects were given brief training with the Polish labels for these sounds and then asked to label the entire stimulus set. This avoids the restrictive English orthographic labels and allows for a test of listeners' abilities to learn distinctions in the stimulus set. Training sets were designed to push learners to either use both the fricative and vowel dimension as the Polish native speaker did, or to use only the fricative dimension which most American English listeners did not do in experiment 1.

Ten subjects (three of whom participated in the labeling described above, subjects 101, 103, 104) were given brief training on specific labels, *sz* and *ś*, and then labeled the entire stimuli set (five repetitions of each stimulus, random order.) Subjects were told that they would be learning two sounds in Polish that are very similar to the English sound *sh*. Training consisted of passive listening to 18 tokens where each token was presented with a simultaneous visual presentation of the desired label. The training procedure continued with a session in which listeners labeled the same tokens with accuracy feedback. In the training phase of the experiment, each token was presented five times for a total of 90 trials (about 7mins); the correct response was presented if a subject responded incorrectly.

5 Two additional Turkish listeners were run through a similar experiment to explore the hypothesis that Turkish listeners are sensitive to fricative variation and ignore formant transitions for this contrast. The results were inconsistent, leaving the issue unresolved. However, one subject during debriefing indicated that he (incorrectly) believed the focus of the experiment was to test his ability to discriminate final unreleased stops and that some of the stimuli (same as this experiment) had final stops.

Subjects were divided into groups differing by the distribution of training stimuli (Figure 5). In one group, five subjects were trained on the 18 tokens closest to the natural tokens (9 alveopalatals and 9 retroflexes), "natural category learners" (NC). The second group consisted of a total of five subjects who were trained on tokens varying maximally on the fricative dimension and minimally on the vowel dimension, "fricative categorizers" (FC). The FC subjects were further divided into two groups, three subjects training on tokens from the alveopalatal end of the vowel continuum, and two from the retroflex end. Training duration was identical for all groups, only the stimulus set for training varied.

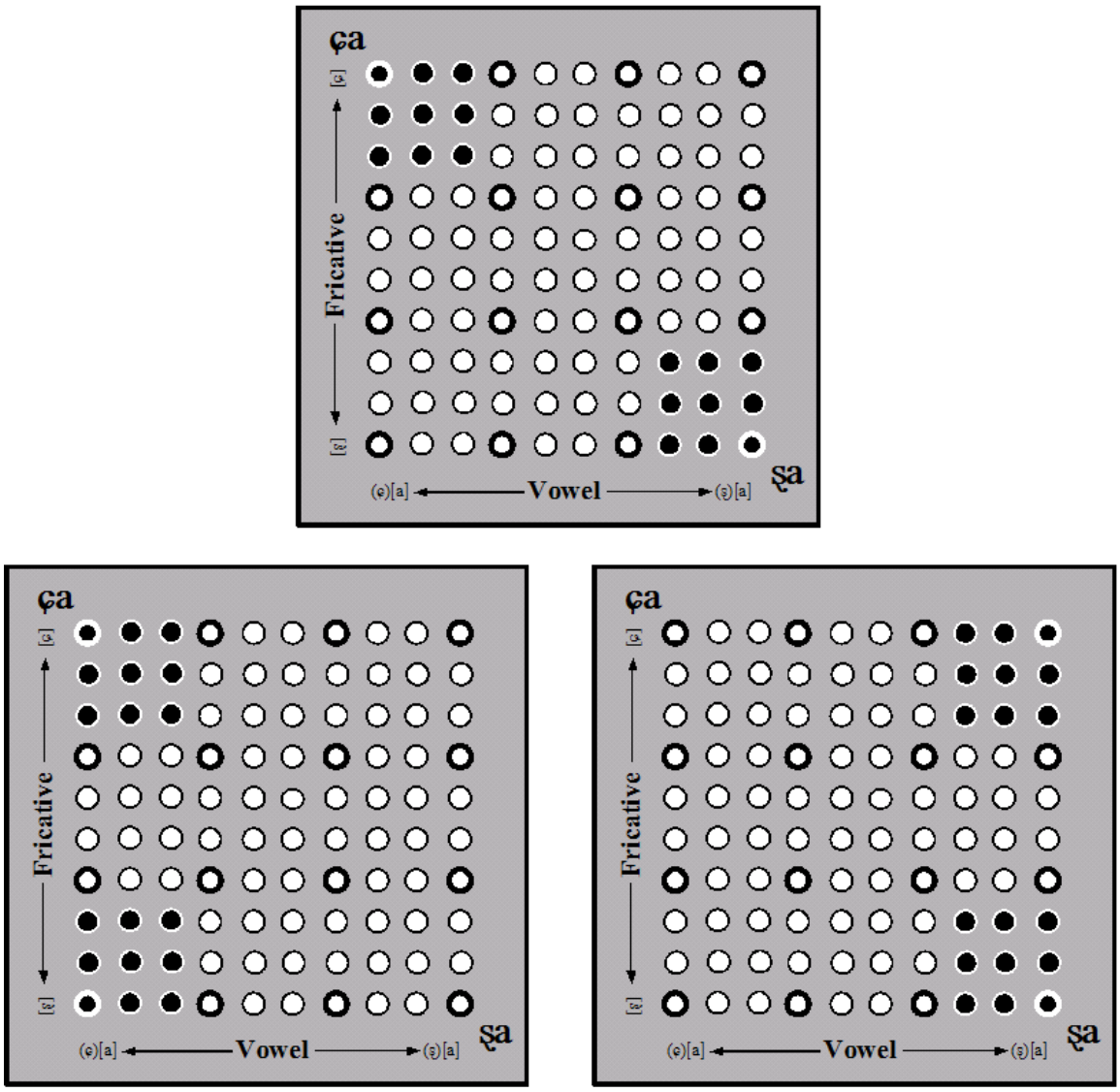

Figure 5: Pilot training sets; dark filled circles represent tokens used for training, white filled tokens were used for testing (along with the training tokens.) Heavier lines indicate tokens used for orthographic labeling. Clockwise from top: 1) natural category group, 2) fricative categorizers (retroflex transitions), 3) fricative categorizers (alveopalatal transitions)

2.3.3 Results

With a single exception, all subjects in the natural category group ignored variation along the fricative dimension and instead relied on the vocalic dimension for determining category membership (see Figure 6). This is in contrast to the Polish listener who used both dimension for categorization. One subject, however, used the fricative dimension only and ignored the vocalic dimension. This is the same subject mentioned above as using *sha* and *ssha* labels and is further unique in being a fluent speaker of Turkish in addition to English. All subjects had a training accuracy greater than 85%.

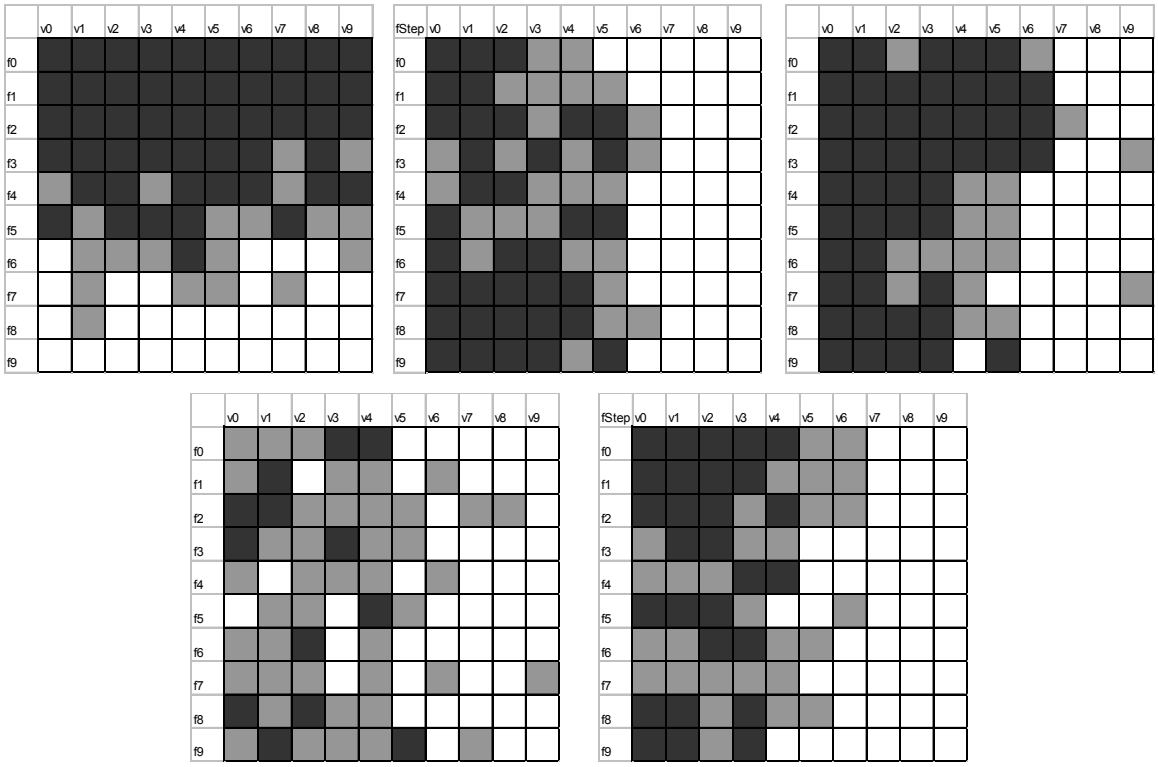

Figure 6: Results from the natural category training subjects. Dark shading indicates the stimulus was labeled as alveopalatal >66%; gray shading indicates between 33% and 66% alveopalatal labeling; white indicates <33% alveopalatal labeling. The top left panel is the subject that categorized along the fricative dimension.

The fricative categorizers show a similar pattern (see Figure 7), despite different training. Only one subject showed overall categorization along the fricative dimension, and a second subject reversed the labels. Performance in training was poorer than that for the natural category subjects and much more variable, from near chance (50%) to 75% accuracy.

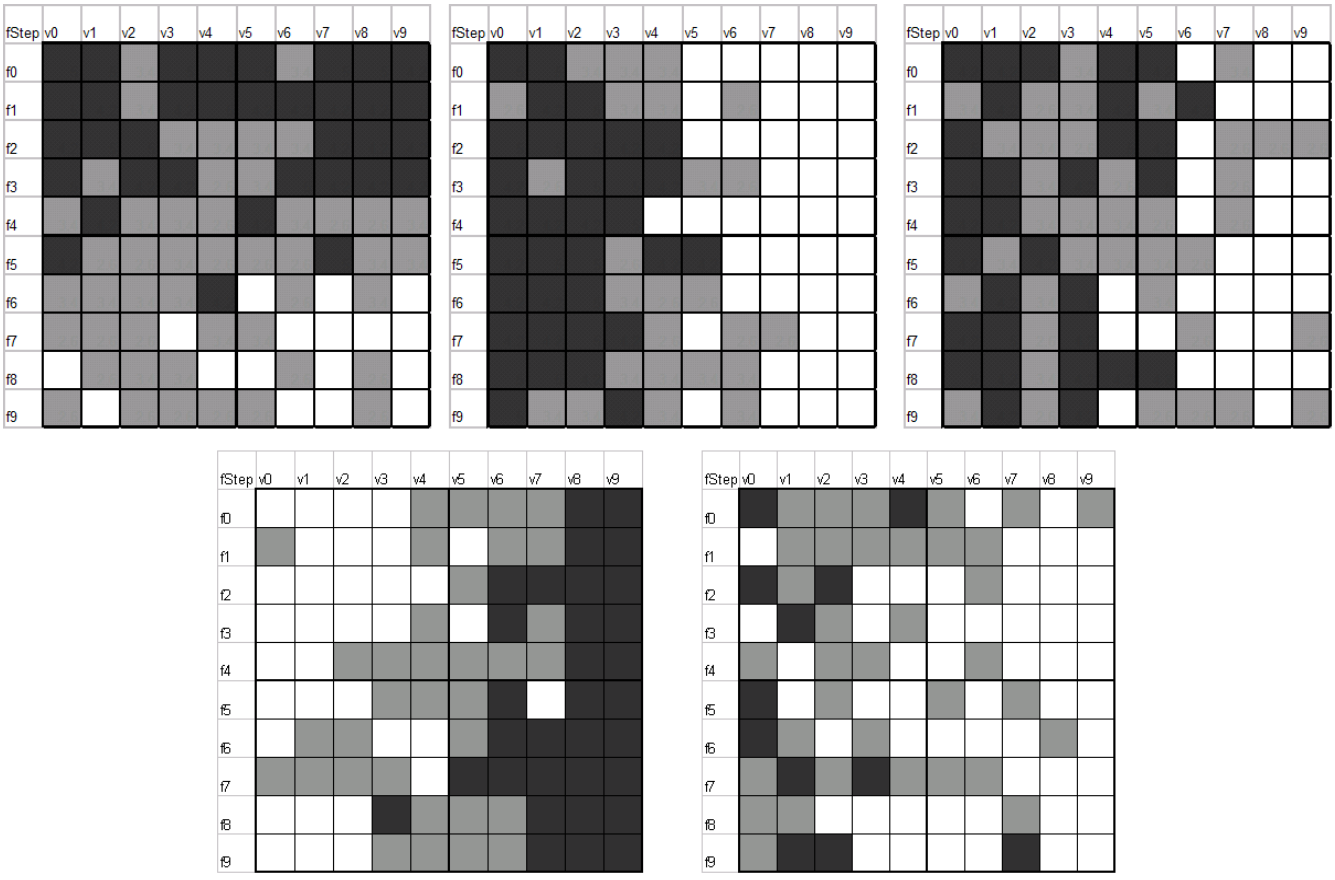

Figure 7: Results of fricative categorizers. Top row shows subjects trained on alveopalatal vocalic tokens and bottom row shows retroflex.

Interestingly, three of these subjects show a change in strategy during the labeling testing, switching from an initial categorization based on the fricative continuum and later switching to one based on the vocalic cues. The subjects who performed better in training showed this pattern. This is most dramatically demonstrated by subject 303, a subject with high accuracy during the training phase from the FC training group (Figure 8).

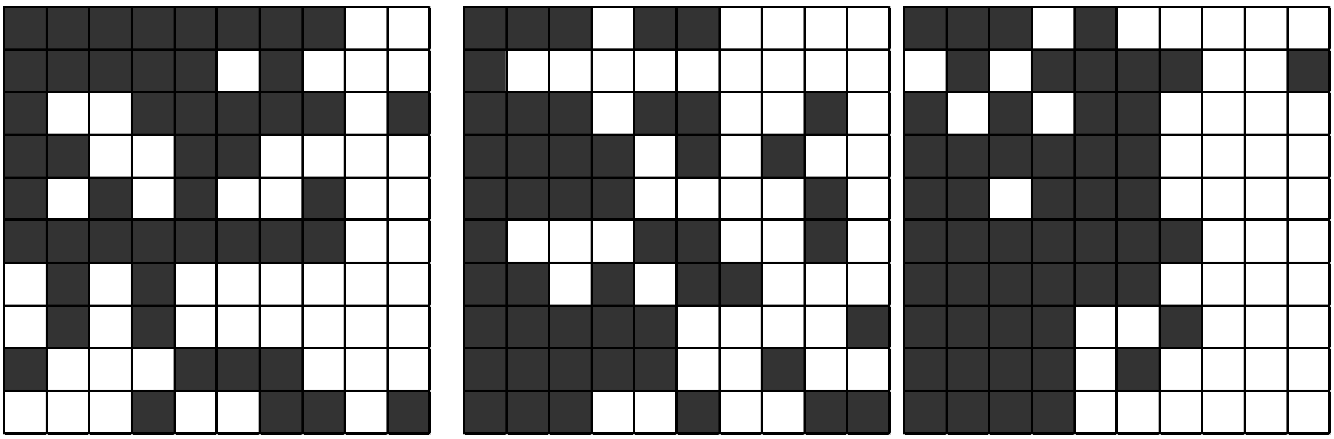

Figure 8: Times series results from subject 303. Left panel represents block 1/5, central panel shows block 3/5, and right panel shows block 5/5. Black indicates stimulus labeled as alveopalatal, white as retroflex.

Overall the results demonstrate that the vocalic cues to the perception of this contrast are more robust than the fricative cues for these subjects and that little training is necessary to achieve reliable categorization along this dimension. With brief training on categorization along the fricative dimension subjects will either ignore those cues outright or gradually shift to using the vocalic cues. However, the fact that some listeners were able to use the fricative dimension during training and over the first block of test indicates that additional training may make the fricative dimension cues more robust and override the vocalic cues.

The role of linguistic experience is tantalizingly hinted at in these results as well. The lone subject with significant experience in another language, Turkish, showed consistently different results from the more monolingual English listeners. Further, none of the listeners used both dimensions to categorize the stimuli as the native Polish speaker did.

2.4 General discussion

In general, these results show that the perceptual differences among the Polish post-alveolars are only partially available to English listeners without training. Although both categories can be perceived by English listeners as /ʃ/ and labeled as such, subjects can reliably perceive a difference between them, labeling the alveopalatal's transition as a palatal glide.

Especially interesting in the results is the ability of English listeners to categorize using one dimension (vocalic) while showing extreme difficulty in using the other dimension (fricative noise). In a training experiment, this allows for a comparison between cues that are easily attended to and those that are difficult to attend to. Further, the initial success some subjects had with training in the fricative dimension suggests that with more extensive training, American English listeners may be able to use this cue reliably.

Also of interest in these results is a contradiction of the Lisker (2001) study. In those experiments English listeners could not reliably identify the alveopalatal and retroflex sibilants as different in the context of full syllables. However, these results show that English listeners can differentiate these sounds, but only using vocalic information. This discrepancy is likely due to differences in experimental design. In this experiment, listeners only had to identify the alveopalatal and retroflex voiceless sibilants as distinct. In Lisker's experiment, subjects also had to identify the Polish dental sibilant, making the three-way place distinction. The subjects were quite reliable in identifying the sibilant as different from alveopalatal and retroflex, which follows Best et al. (2001) as the dental sibilant is quite similar to English /s/ (Lisker 2001, Nowak 2006). It is possible that this additional perceptual demand severely restricted subjects' ability to focus on the relevant distinguishing characteristics of the post-alveolars.

Generally, these stimuli appear to be suitable for a perceptual training study. Training to reliably use fricative noise cue information should be relatively minimal and can be contrasted with vocalic information, which is highly robust. The stimulus set is sufficiently natural with a full specification of information

available in natural speech, yet has variation that can be used to explore changes in perceptual space. Although listeners do not initially attend to the fricative noise difference between Polish [ʂ] and [ɕ], the data here suggest that American English listeners can be trained to use this subtle acoustic cue. The vowel formant difference between Polish [ʂ] and [ɕ] seems to be more salient to American English listeners and thus should form the basis for a strong category if training enforces this tendency.

3.0 References

Boersma, P., and Weenink, D. (2002). PRAAT. Amsterdam, NL: Institute of Phonetic Sciences. University of Amsterdam, NL.

Best, Catherine T., Gerald W. McRoberts, and Elizabeth Goodell (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America, 109 (2)*, 775-794.

Bregman, A.S. (1994). *Auditory Scene Analysis.* MIT Press, Cambridge, MA.

Chistovich, L.A. and Lublinskaya, V.V. (1979). The center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1*, 185-195

Francis, A.L., Baldwin, K., & Nusbaum, H.C. (2000). Effects of training on attention to acoustic cues. *Perception and Psychophysics, 62(8)*, 1668-1680.

Francis, A.L., & Nusbaum, H.C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2), 349-366.

Guion, S.G. & Pederson, E. (2007). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. Munro (Eds.) *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege.* Amsterdam: John Benjamins, 57-77.

Hazan, V. and Barrett, S. (2000). The development of phonemic categorisation in children aged 6 to 12. *Journal of Phonetics 28*, 377-396.

Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages.* Oxford: Blackwell.

Lisker, L. (2001). Hearing the Polish sibilants [s š ś]: phonetic and auditory judgements. In N. Grønnum, & J. Rischel (Eds.), *Travaux du Cercle Linguistique de Copenhague XXXI*. To honour Eli Fischer-Jørgensen (pp. 226–238).Copenhagen: C.A. Reitzel.

Nittrouer (1992). Age-related differences in perceptual effects for formant transistions within syllable and across syllables. *Journal of Phonetics, 20*, 1-32.

Nittrouer (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America. 112(2),* 711-719.

Nittrouer (2006). Children hear the forest. *The Journal of the Acoustical Society of America. 120(4)*, 1799-1802.

Nittrouer and Miller (1997). Predicting developmental shifts in perceptual weighting schemes. The *Journal of the Acoustical Society of America. 101 (4)*, 2253-2266.

Nittrouer and Studdert-Kennedy (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research, 30,* 319-329.

Nowak, P. (2006). The role of vowel transitions and frication noise in the perception of Polish sibilants. *Journal of Phonetics. 34*, 139-152.

Scheffers M.T.M. (1982). The role of pitch in the perceptual separation of simultaneous vowels II. *IPO Annual Progress Report, 17*, 41-45. Institute for Perception Research, Eindhoven.

Zwicker, U.T. (1984). Auditory recognition of diotic and dichotic vowel pairs. *Speech Communication, 3*, 265-277.

Zygis, M. and Hamann, S. (2003). Perceptual and acoustic cues of Polish coronal fricatives. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona: 395-398.

Appendix: Scripts

The Praat script that was used to interpolate the vowels is given below. This script was initially provided on the Praat-users message board and was written by Holger Mitterer. The original can be found here: <http://uk.groups.yahoo.com/group/praat-users/message/1787>. Praat scripts were also used to combine the fricatives, add zeros before and after stimuli to be interpolated, and to concatenate the syllables. These were produced with the assistance of Ronald Sprouse and are available upon request from the author of this paper.

```
#Script for making a continuum from two different vowel signals
#
#
# This script was written by Holger Mitterer, MPI Nijmegen, in order
# to apply the widely used technique of interpolation between two
# natural speech sounds to voiced samples. Usually, mixing two voiced
# sounds gives rise to the (essentially correct) experience of two
# speech sounds, not one ambiguous speech sound. (A possible
# explanation for this phenomenon is grouping by phase.)

#The script uses PSOLA to equate duration and pitch contour,
# and then interpolates between the manipulated sounds

#NOTE: use some zero-padding at the beginning and end
# of the sound to facilitate pitch estimation

# A different method using zero-padding of individual pitch periods
# has been proposed by Stevenson, Hogan, and Rozsypal (1985) in
# Behavior Research Methods, Instruments & Computers, 17(1), 102-106.


#if numberOfSelected("Sound") <> 2
#exit Select 2 Sounds
#endif
Read from file... C:\Stimuli\temp\v_step0_zeroSWS.wav
Read from file... C:\Stimuli\temp\v_step9_zeroSWS.wav

select Sound v_step0_zeroSWS
plus Sound v_step9_zeroSWS
s1$ = selected$("Sound",1)
s2$ = selected$("Sound",2)
```

```
select Sound 's1$'
l1 = Get finishing time
rms1 = Get root-mean-square... 0 0
#get sound length and rms value from Sound 1

select Sound 's2$'
l2 = Get finishing time
rms2 = Get root-mean-square... 0 0
#get sound length and rms value from Sound 2

plus Sound 's1$'
To Manipulation... 0.01 75 300
#cast both sound to Manipulation objects
#with time step of 10 ms and 75 Hz as lower and 300 Hz as upper
#f0 boundary (boundaries determined empirically for the male speaker

for snd from 1 to 2
#for both sounds
s$ = s'snd'$
l = l'snd'
t = ((l1 +l2)*0.5) / l
#variable t is now the mean length of the input sounds in relation to
the sound selected
select Manipulation 's$'
Edit
editor Manipulation 's$'
Add duration point at... 0 't'
Close
endeditor
Get resynthesis (PSOLA)
Rename... temp'snd'
endfor
#the sounds temp1 and temp2 now have exaclty the same length


select Sound temp1
plus Sound temp2
To Manipulation... 0.01 75 300
for snd from 1 to 2
select Manipulation temp'snd'
Extract pitch tier
Rename... p'snd'
endfor
# get the pitch contour of both sounds


select PitchTier p1
Formula... 0.5*(self[col] + PitchTier_p1[col])
# calculate the mean pitch contour
Copy... mean

for snd from 1 to 2
#for both sounds
select Manipulation temp'snd'
plus PitchTier p1
Replace pitch tier
#use the mean pitch contour
```

```
select Manipulation temp'snd'
Get resynthesis (PSOLA)
Rename... org'snd'
endfor

# The two sounds org1 and org2 now have the same length
# and the same pitch contour.


for s from 0 to 10
select Sound org1
Copy... step's'
f = 's'/10
#factor fof the second sound, going from 0 to 1 in steps of 0.1

inv = 1 - 'f'
#factor of the first sound, going from 1 to 0 in steps of 0.1
Formula... self[col] * 'inv'+ Sound_org2[col] * 'f'
nowrms = Get root-mean-square... 0 0
rmsm= 0.5*rms2+ rms1 * 0.5
Formula... self * 'rmsm'/ 'nowrms'
#give sound the mean root-mean-square (rms) of sample values of
bothsounds
# if rms is thought to be a cue in itself, the following alternative
for the calculation of
# the following formula may be used, which also interpolates rms
#rmsm= 'f' * rms2+ rms1 * 'inv'
endfor

select Manipulation 's1$'
plus Manipulation 's2$'
plus Manipulation temp1
plus Manipulation temp2
plus Sound temp1
plus Sound temp2
plus PitchTier p1
plus PitchTier p2
Remove
#clean up the Objects window
```