Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 77 (2016) 84–100

www.elsevier.com/locate/specom

Effects of native language on compensation for coarticulation☆

Shinae Kang[1], Keith Johnson*, Gregory Finley[2]

Department of Linguistics, 1203 Dwinelle Hall, University of California, Berkeley CA 94720-2650, United States

Received 20 May 2015; received in revised form 4 December 2015; accepted 18 December 2015
Available online 29 December 2015

Abstract

This paper investigates whether compensation for coarticulation in speech perception can be mediated by native language. Substantial work has studied compensation as a consequence of aspects of general auditory processing or as a consequence of a perceptual gestural recovery processes. The role of linguistic experience in compensation for coarticulation potentially cross-cuts this controversy and may shed light on the phonetic basis of compensation. In Experiment 1, French and English native listeners identified an initial sound from a set of fricative-vowel syllables on a continuum from [s] to [ʃ] with the vowels [a,u,y]. French speakers are familiar with the round vowel [y], while it is unfamiliar to English speakers. Both groups showed compensation (a shifted 's'/'sh' boundary compared with [a]) for the vowel [u], but only the French-speaking listeners reliably compensated for the vowel [y]. In Experiment 2, 39 American English listeners judged videos in which the audio stimuli of Experiment 1 were used as soundtracks of a face saying [s]V, [ʃ]V, or a visual-blend of the two fricatives. The study found that videos with [ʃ] visual information induced significantly more "ʃ" responses than did those made from visual [s] tokens. However, as in Experiment 1, English-speaking listeners reliably compensated for [u], but not for the unfamiliar vowel [y]. The listeners used visual consonant information for categorization, but did not use visual vowel information for compensation for coarticulation. The results indicate that perceptual compensation for coarticulation is a language specific effect tied to the listener's experience with the conditioning phonetic environment.

Keywords: Speech perception; Compensation for coarticulation; Linguistic experience; Direct realism; Audiovisual perception.

1. Introduction

1.1. Three modes of speech perception

General properties of the auditory system determine what can and cannot be heard, what speech cues will be recoverable in particular segmental contexts, and to at least some extent how adjacent sounds will influence each other. For example, the cochlea's nonlinear frequency scale probably underlies the fact that no language distinguishes fricatives on the basis of frequency components above 6000 Hz (Johnson, 2012). Similarly, limitations on the auditory system's ability to detect the simultaneous onset of tones at different frequencies probably underlies the fact that the most common VOT boundary across languages is at about ± 30 ms (Pastore and Farrington, 1996).

In addition to these general auditory factors, speech perception may also be shaped by phonetic knowledge. Because language users are both speakers and listeners, we come to the task of speech perception with a base of knowledge that makes available a "phonetic mode" of listening (or "speech mode"; Liberman and Mattingly, 1985). [Strictly speaking, Liberman and Mattingly, 1985 "speech mode" is not completely synonymous with our concept of the "phonetic mode" because we use the term "phonetic mode" in a more general sense to contrast knowledge-based phonetic processing with general auditory processing.] By hypothesis, the phonetic mode elaborates and reinterprets the auditory image of speech. Thus, the phonetic mode may underlie the tendency for multimodal information to be combined into a phonetic

☆ The original version of this paper was selected as one of the best papers from Interspeech 2010. It is presented here in revised form following additional peer review.

* Corresponding author. Tel.: +1 510 642 2757; fax: +510 643 5688.
E-mail addresses: sakang2@berkeley.edu (S. Kang), keithjohnson@berkeley.edu (K. Johnson), finley@berkeley.edu (G. Finley).
[1] Present address: Georgetown University, Washington, DC.
[2] Present address: University of Minnesota, Minneapolis, MN.

http://dx.doi.org/10.1016/j.specom.2015.12.005
0167-6393/

percept (McGurk and Macdonald, 1976), and may explain how the perception of sine wave analogs of speech may suddenly shift from nonphonetic to phonetic (e.g. Remez et al., 1981). Additionally, the phonetic mode of speech perception is probably also involved in the perceptual coherence of signal components that might not ordinarily be grouped with each other in the phenomenon of duplex perception (Bregman, 1990; Whalen and Liberman, 1987) or in the integration of asynchronous audio signals (Nygaard and Eimas, 1990).

Scholars differ in their view of whether the phonetic mode of perception is innate or learned. For example, according to Liberman and Mattingly (1985), the speech mode of listening is innate and does not require experience as a speaker. On the other hand, Best (1995) assumed that the phonetic mode is grounded in experience-based perceptual learning and this underlies the strong tendency to hear foreign speech in terms of native segments. Fowler (1986, 1996) places much less emphasis on learning and in this respect is similar to Liberman and Mattingly's view.

Speech perception is also shaped by lexical knowledge. The fact that the listener's ultimate aim in speech communication is to figure out what words the speaker is saying underlies lexical effects in speech perception. For example, perceptual errors ("slips of the ear") overwhelmingly result in words (Bond, 2005). Similarly, Ganong (1980) showed a lexical effect on phoneme identification. In a "tash-dash" VOT continuum there are more "d"-responses, consistent with the word "dash", than in a "task-dask" continuum. Similarly, a missing or obliterated phoneme can be perceptually restored (Pitt and Samuel, 1995), and the restored phones interact with phonetic mode processes like compensation for coarticulation (Elman and McClelland, 1988; but see McQueen et al., 2009).

Researchers who primarily focus on one or the other of these three aspects of speech perception (auditory, phonetic, or lexical) are often critical of the others (e.g. Fowler, 2006 against the exclusive effects of auditory spectral contrast on compensation for coarticulation; McQueen (2006) against direct lexical involvement in speech perception; and Diehl and Walsh, 1989; Lotto and Kluender, 1998 against a specifically phonetic mode of processing). Our view is that it is more plausible to assume that all three factors are simultaneously involved in speech perception. Indeed, recent findings from neuroscience (cf. Hickok and Poeppel, 2004) indicate that all three are simultaneously involved in speech perception. Ultimately, a successful theory of speech perception has to predict which listening circumstances will engage greater or lesser reliance on phonetic processing, or lexical processing, and what aspects of speech perception ultimately derive more from auditory processing than from specifically linguistic processing.

1.2. Compensation for coarticulation

In this paper, we explore how the phonetic mode of listening may be shaped by linguistic experience in a compensation for coarticulation task. Our experiments on compensation do not test for auditory contrast or lexical activation effects, but we are aware of the literature in these areas. For example, in the literature on whether a lexically biased percept can induce compensation for coarticulation (Elman and McClelland, 1988; Pitt and McQueen, 1998), compensation is assumed to exist as a separate, phonetic mode, phenomenon that can be used as a diagnostic to determine whether the restored phoneme is truly restored. We do not go further in lexically induced compensation for it is beyond the scope of this study.

Compensation for coarticulation (Mann, 1980; Mann and Repp, 1981) is a listener's perceptual "demodulation" of coarticulatory information during speech perception. For example, Mann and Repp (1981) found that the lower fricative pole induced by adjacent vowel lip rounding in [s] did not induce the percept of a more alveopalatal fricative [ʃ], while the same fricative noise paired with the unrounded vowel [a] does sound more like [ʃ]. This phenomenon of attributing one aspect of the acoustic signal (lower pole frequency) to coarticulation with a neighboring vowel, and thus not only an inherent property of the fricative itself, is a prototypical case of compensation for coarticulation. Compensation has been investigated in many studies of consonant–vowel interactions in consonant place perception (e.g. Mann and Repp, 1981; Mitterer, 2006; Smits, 2001; Whalen, 1981), vowel perception (Holt et al., 2000), and consonant voicing perception (Diehl and Walsh, 1989), as well as in vowel–vowel interactions (Fowler and Smith, 1986; Bradlow and Bent, 2002), and in consonant–consonant interactions (Fowler, 2006; Lotto and Kluender, 1998; Mann and Repp, 1981; Pitt and McQueen, 1998).

Much of this literature is steeped in controversy regarding the basis of the compensation mechanism—whether it is due to the auditory interaction between adjacent segments, or due to a phonetic mode of processing "undoing" the gestural interactions inherent in speaking and thus an indication that speech is perceived in terms of phonetic gestures. While some researchers have suggested that auditory spectral contrast plays a primary role in the phenomenon of compensation for coarticulation (Johnson, 2011; Lotto and Kluender, 1998), several studies have provided evidence showing that spectral contrast alone cannot capture the whole phenomenon (e.g. Fowler, 2006).

For example, Mitterer (2006) found an effect of visible lip rounding by Dutch listeners and concluded that compensation for coarticulation has a phonological basis. He studied perception of a [si]-[sy] fricative continuum, first testing whether compensation for vowel rounding can be replicated with non-speech audio that imitates critical acoustic characteristics (spectra contrast etc.), and second testing whether compensation for vowel rounding (in natural speech tokens) increased when the participants saw audio/visual stimuli with lip rounding during the vowel. The participants showed no compensation effect for the non-speech audio, and an increased effect for AV stimuli. Based on these results, he concluded that the basis of compensation for coarticulation is not solely auditory.

Similarly, Viswanathan et al. (2010) found evidence that phonetic knowledge impacts compensation for coarticulation. They tested compensation in stimuli like the [aɹda]-[alga] continuum used by Mann (1980). Mann found that a greater number of "d" responses on a [da]-[ga] continuum with the precursor syllable [aɹ] than with the precursor [al]. Mann attributed this to a compensation for coarticulation between [ɹ] or [l] and the following stop. With a more back tongue-tip articulation, [ɹ] context (presumably) causes a backer [d] closure location. Listeners' behavior in the perception test is thus a 'compensation' for this coarticulation because they allow backer consonants to still be called "d". The key component of the argument is that English [ɹ] produces a more backed [d] than [l] does. Viswanathan et al. (2010) replicated Mann (1980) with a set of four precursor segments. In addition to American English precursors like those used by Mann ([al] and [aɹ]) they also included two precursors with the Tamil liquids [r] (an alveolar trill) and [ɭ] (a retroflex lateral). This design gives a matrix with two front segments – AE [l] and Tamil [r] – and two back segments – AE [ɹ] and Tamil [ɭ]. Of these segments only [l] has a high F3 value, so if the compensation effect is due to auditory interaction of F3 in the precursor with F3 of the [da]/[ga] continuum (as suggested by Diehl et al., 2004) then the pattern of results should be different from the pattern predicted by a theory based on the backness of the consonant articulation (Fowler, 1996). The results of the experiment suggested that the articulation of the segment, not its F3 value, is what mattered. The identification curves for precursors [al] and [ar] were practically identical to each other and differed from the curves for [aɹ] and [aɭ]. This result suggests that the actual place of articulation of the unfamiliar sounds (for the English-speaking listeners in their study) trilled [r] and retroflex [ɭ] were recovered accurately during perception, despite the fact that these listeners had no personal experience with them. Viswanathan et al. (2010) concluded from this that the auditory account alone is not sufficient in explaining compensation for coarticulation. Taken together, the investigation of how compensation for coarticulation in audiovisual modality is different from that in audio-only modality allows us to address if compensation for coarticulation, the perceptual phenomenon which largely engages a phonetic mode of listening, has an articulatory basis.

1.3. Cross-linguistic studies

In the present study, we aim to explicitly test if articulatory phonetic knowledge extends to foreign sounds with which listeners have little experience. The purpose of using the non-native sounds is to address whether the gestural knowledge that listeners rely on during speech perception is also affected by a linguistic factor. Although the role of experience has been extensively studied in various topics such as phonetic categorization during first language acquisition (e.g. Kuhl et al., 1992) and non-native sound perception (Best et al., 1988), relatively few studies of cross-linguistic

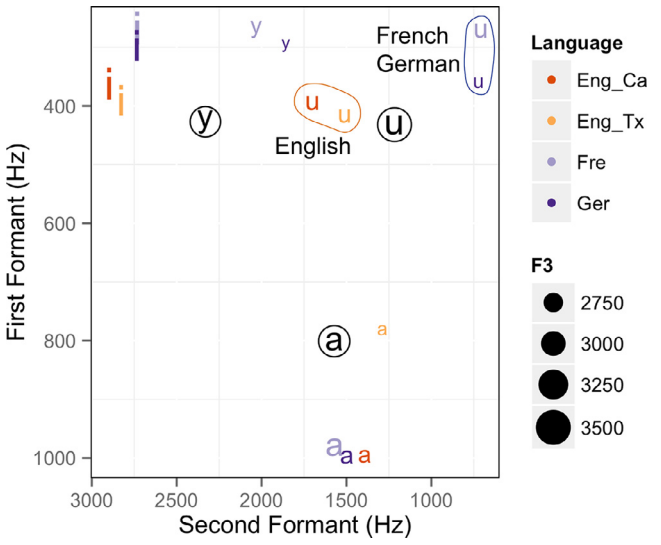

Fig. 1. The first three vowel formants of [i], [u], [y] and [a] for female speakers of German (Strange et al., 2007), French (Ménard et al., 2007), and two varieties of English (Texas: Yang, 1996; California: Hagiwara, 1995). The value of F3 is indicated by the size of the vowel symbol. Formant values of the stimuli used in Experiments 1 and 2 are shown with the circled [i], [a], and [u] symbols (the size of which is also tied to the value of F3).

compensation for coarticulation have been reported, and there are some inconsistencies in the literature.

One of the most important cross-linguistic studies of compensation for coarticulation in the literature found no difference between speakers of different languages. Mann (1986) compared Japanese-speaking and American English-speaking listeners' responses to a [da]-[ga] continuum in the context of [al] and [aɹ] precursor syllables (extending Mann, 1980) and concluded that results suggest that there is a language-independent phonetic mechanism because a compensation effect was found for both groups of listeners. This is very interesting because Japanese lacks the /l/-/ɹ/ distinction found in English. This result was taken to suggest that compensation is a property of human speech perception regardless of one's linguistic experience with the particular speech sounds involved. Mann concluded that "there exists a universal level where speech perception corresponds more objectively to the articulatory gestures that give rise to the speech signal" (p. 192). Although it is remarkable that segments that couldn't be reliably identified as different by Japanese listeners nonetheless produced different compensation effects, Mann's conclusion may have been premature because the effects appear to be different for the two groups of listeners. Mann did not report any statistical tests comparing context ID curves to the baseline ID curves, but the pattern is quite striking when you plot the baseline and context conditions from her Fig. 1 on the same graph for each group of listeners. Japanese listeners showed a shift from baseline (no context) with the [aɹ] precursor, while English listeners showed the opposite shift from baseline with the [al] precursor. Presumably, some aspect of prior linguistic experience was responsible for the difference

between Japanese-speaking and English-speaking listeners in this study.

As noted above, Viswanathan et al. (2010) found little or no effect for language-specific experience. In their study, American English listeners were able to compensate for non-native Tamil liquids despite their lack of experience in perceiving the Tamil lateral or retroflex sounds. More interestingly, the listeners' pattern of compensation matched the articulation for the sounds (front versus back liquids) and not their acoustic F3 frequency. Fowler et al. (1990) can also be taken as evidence that compensation for coarticulation is not tied strictly to linguistic knowledge. They found a compensatory-like effect in which /l/ and /r/ influenced the perception of /da-ga/ stimuli with young infants who have little linguistic experience.

In contrast to these studies, several other researchers have found that language-specific experience does influence the degree of compensation. The studies seem to suggest that the compensation effect is not a result of a "universal level" in speech perception. For example, Beddor and Krakow (1999) found that perceptual compensation for vowel nasalization differed for Thai and American English speakers in that Thai listeners showed a smaller perceptual compensation effect which mirrored the tendency for Thai to show less coarticulatory nasalization on vowels in CVN sequences than found in English. Similarly, Beddor et al. (2002) investigated the perceptual response to vowel-to-vowel coarticulation. In their acoustic/phonetic study, Shona and English were found to have different coarticulatory patterns, and their perception study showed that Shona and English listeners' compensation patterns largely matched their production differences. Finally, Harrington et al. (2008) found that the perceptual compensation for the coarticulation of /u/ in a fronting context (before the alveolar /d/) was smaller for younger speakers of British English who had a smaller degree of coarticulation (due to /u/ fronting) than for older speakers who had a larger coarticulatory context effect. This research suggests that compensation for coarticulation has a learned component.

We report two studies in this paper. The first is a cross-linguistic study that is designed to test for the language specificity of compensation for coarticulation by comparing compensation for vowel rounding by English-speaking and French-speaking listeners with the vowels /u/ and /y/ – a contrast which is found in French but not in English. The second study is an audio–visual speech perception experiment with English-speaking listeners in which we show the listeners just how rounded the lips are in these particular tokens of /u/ and /y/ in a test of their ability to pick up phonetic information from the audio-visual signal. The first experiment tests for a learned, language specific component of speech perception, and the second seeks evidence of a language independent phonetic mode in compensation for coarticulation. To preview our results, we found strong evidence of language specificity, but we also found that this language-specific pattern of perception persists despite conflicting visual information.

2. Experiment 1

Experiment 1 is a cross-linguistic study of compensation for coarticulation in which we test for a linguistic component of the phenomenon by comparing compensation effects for native and nonnative sounds. Smits (2001) used the high front round vowel [y] in a compensation study with Dutch-speaking listeners, while Mann and Repp (1981) used the high back round vowel [u] with English-speaking listeners. The present experiment uses both of these round vowels as potential triggers of compensation. One group of listeners (French) have native language familiarity with both [y] and [u], while the other group (English) is only familiar with [u] (albeit a more front [u] than the one in French). The phonetic details of [y] and [u] in French, English, and German set the stage for our cross-linguistic study of speech perception.

The vowel broadly transcribed as /u/ is produced with a lower F2 in French than in English. Fig. 1 shows this with data drawn from female speakers in four different studies of vowel acoustics. The "[u]" of American English is acoustically closer to the acoustic values of [y] in French and German. Not surprisingly both [u] and [y] of French and German tend to be identified as "u" by speakers of American English (Strange et al., 2009), but interestingly this tendency is stronger for [u] than it is for [y]. Strange et al. (2009) found that German [u:] was identified as "u" 94% of the time, while German [y:] was labeled "u" only 77% of the time. French [u] was labeled "u" 84% of the time, and [y] was labeled "u" only 52% of the time. So despite the acoustic similarity of English [u] and French [y] in the acoustic vowel space, listeners don't always hear [y] as "u". F3 frequency may play a role in this pattern of identification, but also there may be a tendency for American English listeners to expect [u] vowels to have a lower F2 than it actually does - a perceptual hyperspace effect (Johnson et al., 1993; Johnson, 2000).

Articulatorily, there is evidence that in British English the higher F2 of [u] is due to tongue fronting (Harrington et al., 2011), and given the sensitivity of [u] F2 to coronal consonant context among speakers in California (Kataoka, 2011) we suppose that our speakers also produce [u] with a fronted tongue, compared with German or French. We also have some evidence regarding the degree of lip rounding during [u] and [y] in French, German, and English (see Fig. 2). The data in Fig. 2 are of the horizontal and vertical dimensions of lip opening at the acoustic midpoint of the vowels. Noiray et al. (2011) found that these dimensions are good for measuring labial coarticulation patterns in French and English. Linker (1982) also found that lip protrusion is a useful measure for at least some languages. Data for French and Cantonese lip positions in Fig. 2 and a portion of the American English data come from Linker (1982), and the remainder of the data come from our own measurements of speakers for our experiments (the German data) and from a couple of participants in other studies in our lab (Johnson and Bakst, 2014). As with the vowel formant data above, these data are from women. The key observation is that lip rounding for English [u] is not as extreme as it is for French or German.

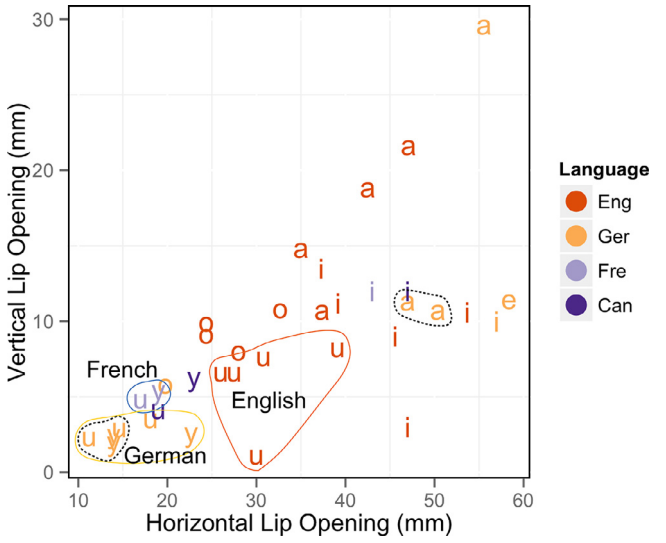

Fig. 2. Horizontal and vertical opening of the lips for vowels in four different languages. The German data are from two speakers, one of whom was the speaker for the stimuli used in the experiments reported here (circled with dashed lines). The English data are fromboth Linker (1982) and from our own recordings in a separate project. The English /u/ and /i/ with very small vertical opening are from Linker (1982). The French and Cantonese data are from Linker (1982).

Benguerel and Cowan (1974) found that both [u] and [y] induce rounding coarticulation in consonants preceding the vowel. Not only is there measurable coarticulation, the presence of rounding on excised consonants is perceptible enough to give listeners an edge in guessing the following vowel (Benguerel and Adelman, 1976). Noiray et al. (2011) criticized the speculation offered by Byrd and Saltzman (2003) that French and English have different types of anticipatory coarticulation ("look ahead" coarticulation in the case of French and "time-locked" coarticulation in the case of English) and presented data suggesting that the patterns of coarticulation seen in both language is consistent with a "movement expansion model" of coarticulation. Noiray et al. found that English and French showed anticipatory rounding coarticulation with comparable kinematic profiles.

The experiment that we conducted builds on these phonetic observations in a study of the language-specificity of compensation for coarticulation. English-speaking listeners are not familiar with the contrast between [y] and [u] that is found in French and German, have a tendency to be less sure about what vowel is actually intended when they hear [y], and produce in their own speech much less lip rounding in [u] than is produced in either German for French. None of these language-specific phonetic details should matter if compensation is driven by the articulatory reality of the speech being presented to the listener – a prediction in line with Fowler's (1996) theory of speech perception that seems to be supported by Viswanathan et al.'s (2010) results. Our hypothesis, on the contrary, is that linguistic experience will matter for compensation for coarticulation.

It has been found that phonetically trained listeners are not very good at detecting vowel rounding for unfamiliar vowels from acoustic signals alone (Lisker and Rossi, 1992; Traunmüller and Öhrström, 2007). Lisker and Rossi (1992) tested whether or not French-speaking participants could identify the rounding of each vowel in an audio-only, in a visual-only, and in AV-congruent and AV-incongruent conditions. The participants' summed responses indicate that seeing the face significantly affected their rounding judgments, even when they were prompted to primarily depend on the auditory signal. Traunmüller and Öhrström (2007) investigated how Swedish vowels /i/, /y/, /e/ and /ø/ are identified by Swedish speakers in a cross-dubbed audiovisual modality with incongruent cues to vowel openness, roundedness, or both. The results showed that identification of vowel height is based primarily on the audio signal, whereas the identification of vowel rounding is based mainly on the visual signal.

On the other hand, Ettlinger and Johnson (2009) found that experience with a feature such as rounding did not translate to skill in dealing with the same feature on unfamiliar sounds. They measured the perceived similarity of a set of German vowels for listeners whose native languages were English, French and Turkish. The vowels were [i], [ɪ], [y], and [ʏ]. These differ by rounding and tenseness. The fact that a front rounding contrast is not present in English while the tense/lax contrast is not present in French or Turkish yielded different predictions as to whether having a sound in the inventory vs. knowing a featural contrast would play a more important role in similarity judgments. The experiment showed that French and Turkish listeners rated the [i]–[y] pair as more distinct than did the English listeners, while English listeners judged [i]–[ɪ] as more distinct than did the French/Turkish listeners. Interestingly, English listeners found [y]–[ʏ] pair less distinct than the other two groups, showing that knowing a featural contrast (tense/lax) did not extend to a pair of unfamiliar sounds.

These observations lead us to suspect that compensation for rounding coarticulation in fricative perception may depend on the listener's familiarity with the specific [+ round] vowels used in the experiment. If linguistic experience guides compensation for coarticulation, then we expect that American English-speaking listeners will show less compensation for rounding coarticulation than will French-speaking listeners.

2.1. Methods

2.1.1. Subjects

Forty-two listeners between ages 19 and 27 participated in the experiment. Twenty-one participants were native speakers of American English who were attending University of California, Berkeley at the time of participation. The other 21 participants were native speakers of French who were recruited at Université Pierre-Mendès-France, Grenoble, France. None of the subjects reported any speech or hearing problems. Several participants in the American English group were bilinguals or equally fluent in other languages including Hindi, Spanish, and Farsi, but none of them was a native speaker of any language with front rounded vowels. Also, it is likely (but

Table 1
Formant frequency measurements of the vowel portions of the stimuli used in Experiments 1 and 2. Measurements are from the onset of the vowel and from the temporal mid-point of the vowel.

		Vowel onset	Vowel mid-point
[a]	F1	662	801
	F2	1973	1572
	F3	3041	3087
[u]	F1	374	432
	F2	1678	1217
	F3	2975	3208
[y]	F1	305	427
	F2	1935	2331
	F3	2999	3046

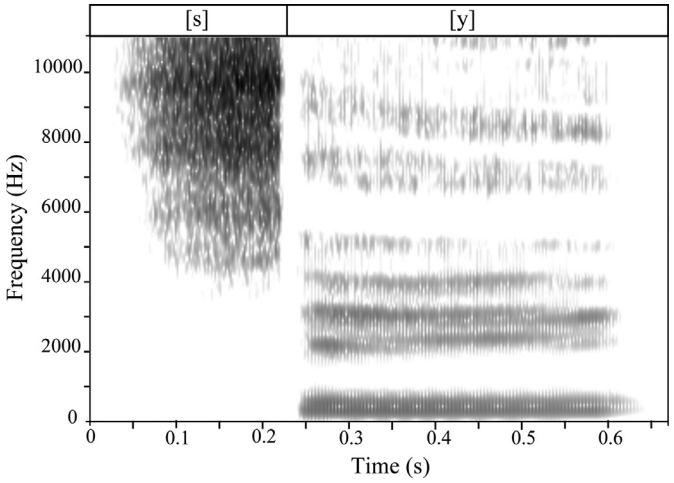

Fig. 3. A spectrogram of an experimental stimulus – the [s] endpoint (Token 1) concatenated with the natural vowel [y].

the relevant data was not collected at the time) that at least some of the participants in the French group were bilinguals in German and/or English.

2.1.2. Stimuli

Six CV syllables ([sa], [su], [sy], [ʃa], [ʃu], [ʃy]) consisting of a fricative (/s/ or /ʃ/) and a vowel (/a/, /u/, or /y/) were first recorded by a female native speaker of German with a Canon Model XF 100a camcorder with high definition audio (16 bit uncompressed sampling rate of 44100 Hz) and video (740 × 480 pixels per frame, 30 frames/s). German was used as the stimulus language because we wanted neither the English-speaking nor the French-speaking listeners to have a "native language advantage" with the stimuli (Bradlow and Bent, 2002). They were slightly foreign to both sets of listeners. The vowels and consonants in these stimuli are part of native phoneme inventory of German.

The vowels from selected, representative tokens of [sa], [su] and [sy] were segmented from the audio track and saved as separate .wav files. The onset of voicing was considered to be the beginning of the vowel for this purpose. To prevent audible editing artifacts the vowels were given a brief (50 ms) linear fade-in. Table 1 shows acoustic vowel formant measurements at both the vowel onset and at the midpoint of the vowel. The midpoint measurements were also shown above in Fig. 1.

Endpoint [s] and [ʃ] tokens were synthesized using the Klatt terminal analog synthesizer (Klatt, 1980) modeled after the naturally produced [s] and [ʃ] preceding vowel [a] (where the fricatives are spectrally most different from each other). The synthetic fricatives were 240 ms in length and were adjusted so that their amplitude relative to the vowels matched the relative amplitude of the natural fricatives. The synthesized fricatives and the extracted natural vowels were then concatenated to produce CV syllables. Fig. 3 shows an example.

The synthetic [s] and [ʃ] were used as the endpoints of a 9-step synthetic fricative continuum in which the formant frequency and amplitude parameters stepped linearly from values for [s] to values for [ʃ] (see Table 2). The frequency steps of the continuum were equally spaced on the bark frequency

Table 2
Synthesis parameters and acoustic analysis of the synthetic fricatives used in Experiments 1 and 2. The first three sections of the table (overall gain, formant amplitudes, formant frequencies) show values of the varying parameters that were used in the Klatt formant synthesizer. The last two sections of the table show the results of a spectral moments analysis and a one-pole LPC acoustic analysis of the synthetic fricative noises.

			[s]								[ʃ]
		Step	1	2	3	4	5	6	7	8	9
Synthesis Parameters	Overall gain	g0	66	64	62	61	59	57	56	54	53
	Formant Amplitudes	A3	35	38	42	46	50	53	57	61	65
		A4	44	47	51	54	58	61	65	68	72
		A5	58	60	62	64	67	69	71	73	76
		A6	53	55	57	55	62	64	66	68	71
	Formant Frequencies	F3	4661	4341	4042	3764	3504	3262	3036	2825	2628
		F4	5875	5775	5677	5581	5487	5394	5303	5213	5125
		F5	7812	7661	7514	7369	7227	7088	6952	6818	6687
		F6	9625	9343	9062	8781	8500	8218	7937	7656	7375
Acoustic Measurements	Moments Analysis	COG	9272	8776	8294	7946	7575	6967	6087	5468	4693
		SD	1089	1177	1265	1344	1524	1754	1993	1973	1985
		Skew	−0.74	−0.82	−0.73	−0.89	−0.61	−0.61	−0.17	0.14	0.81
		Kurtosis	6.96	3.25	2.49	3.23	4.04	1.29	−0.43	−0.87	1.16
	LPC	Pole	9149	8861	8510	8223	7953	7504	6931	6294	5731

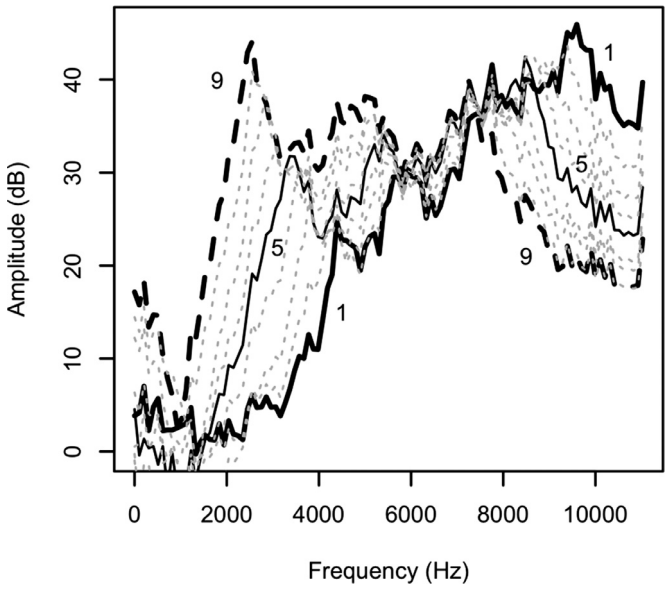

Fig. 4. LTA spectral slices of the nine synthesized fricatives. The spectrum of token 1 is drawn with a heavy solid line, the spectrum of token 9 is drawn with a heavy dashed line, and the spectrum of token 5 is drawn with a light solid line.

scale. Concatenating the continuum fricatives with each of the three vowel environments ([a] from [sa], [u] from [su], and [y] from [sy]) resulted in the 27 stimulus tokens used in this experiment. Fig. 4 shows the average spectrum for each of the synthetic fricative tokens.

Table 2 shows the synthesis parameters for the fricative continuum as well as acoustic measurements taken using moments analysis (Forrest et al., 1988) at the fricative midpoint and an LPC analysis near the fricative offset.

2.1.3. Procedure

The 27 CV tokens (9-step continuum × 3 vowels) were iterated seven times and the resulting list of 218 trials was presented to the participants in random order. The order was randomized separately for each listener. The participants heard one CV-stimulus at a time over headphones (AKG K240 Studio Headphones) and were asked to identify the initial fricative as either 's' or 'sh'. The labels "s" and "sh" were printed on a computer screen in front of the subject ("s" on the left side of the screen and "sh" on the right) and the subject entered a response by pressing either the "1" (for "s") or "0" (for "sh)" key on a standard computer keyboard. The intertrial interval was 1 s.

2.1.4. Statistical analysis

In order to test for language-specificity in perceptual compensation for coarticulation, the responses were analyzed in two ways – first using mixed effects logistic regression predicting the raw 's' versus 'sh' response data, and second using a repeated measures analysis of variance of the calculated category boundaries for each listener in each experimental condition.

The mixed effects logistic regression model had three predictors. In the best-fitting model, TOKEN (range: 1–9) was treated as a continuous variable and both linear and cubic terms were entered into the model. Treating token as continuous variable made it possible to build models maximal models (Barr et al., 2013). A model treating token as an ordered categorical effect showed that the linear and cubic terms were significant, but the maximal model failed to converge probably because of the large size of random effects strucure ($42 + 7 * 3 * 42$ random effects to estimate). It was desirable to fit models with TOKEN by VOWEL random slopes to control for by-subject random variability in the effect of vowel on the identification function. The other factors were more straight-forward. VOWEL (/a/, /u/, /y/) indexed the context vowel – the default value was /a/. LANGUAGE (English vs. French) indexed the native language of each listener – the default category was "English". Finally, LISTENERS (42 levels) was treated as a random factor, and we entered random slopes by LISTENER for the VOWEL by TOKEN interaction (the random term was (TOKEN * VOWEL | LISTENER). The dependent variable was listeners' RESPONSE ('s' vs. 'sh'). 's'-responses were coded as 1 and 'sh'-responses as 0, thus positive model coefficients indicate greater probability of 's'-responses and negative coefficients indicate greater probability of 'sh'-responses. Likelihood ratio tests were used to select the random structure of the model (Pinheiro and Bates, 2004), and the fixed effect coefficients were evaluated by t-test.

The second analysis was based on calculated category boundaries. For each subject in each vowel environment, we fit a four parameter "Gompertz" logistic curve to the proportion 's' responses identification function. One parameter of this curve fit is the 50% cross-over location in the identification curve. The Gompertz fit is a useful function in this context because it does not require symmetry of shape over the two halves of the curve and allows curves to fail to reach $p = 0.0$ or $p = 1.0$ at the endpoints of the curve. When the curve could not be fit to the data (less than 5% of the cases) the boundary was estimated as the sum of the proportion of 's' responses over the entire continuum plus 0.5. We analyzed these data with a repeated-measures Analysis of Variance and with a set of planned comparison *t*-tests.

2.2. Results

All of the listeners showed relatively high accuracy on the endpoint tokens #1 and #9 for all three vowels (all were above 83% correct). The overall proportions of 's'-responses as a function of fricative token number in the three vowel conditions by English and French listeners are shown in Fig. 5.

The pattern in Fig. 5 reflects how the auditory continua were made. Since there were nine tokens created by interpolating the synthesis parameters of the two endpoint fricatives, the proportion of 's'-responses naturally decreases along the continua from Token 1 to Token 9. Round vowels elicited more 's' responses for both English and French listeners (the 50% cross-over boundary for [u] is at a higher stimulus number than it is for [a]), indicating that listeners compensated

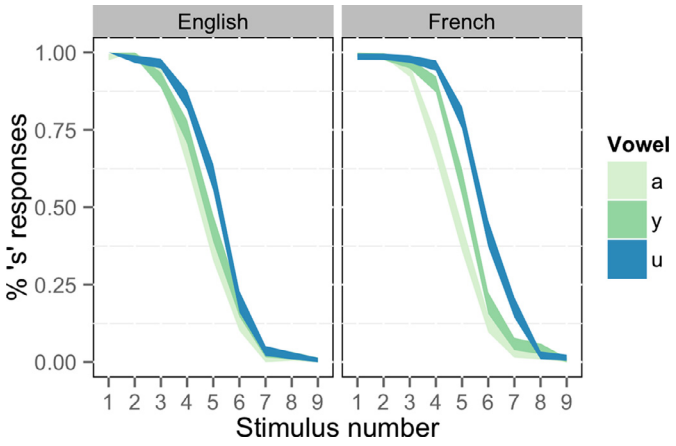

Fig. 5. Results of Experiment 1. Identification curves showing proportion of 's'-response averaged across all speakers in three vowel environment by English Listeners (Left) and French Listeners (Right).

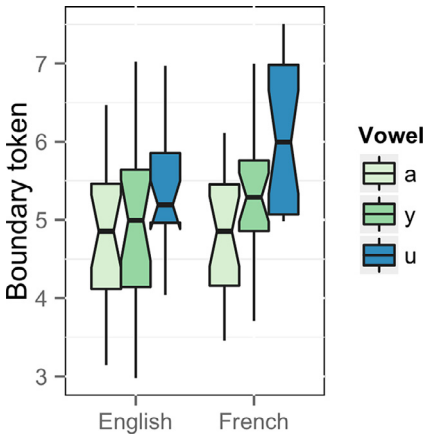

Fig. 6. Distributions of the boundary locations for [a], [u], and [y] for the English-speaking and French-speaking listeners.

Table 3
Experiment 1 model selection. Models 1–4 have the random effects definition: (1 [+ Token ∗ Vowel | Subject). The χ^2, degrees of freedom, and *p*-Values in this table are for likelihood ratio tests as discussed in the text.

Model	AIC	χ^2	df	*p*-Value
1 Token	3089	–	–	–
2 Token ∗ Vowel	3046	50	4	< 0.001
3a Token ∗ Vowel ∗ Language	3047	11.06	6	= 0.07
3b Token ∗ Vowel + Vowel:Language	3042	10.9	3	= 0.012
4a Token^3 + Token ∗ Vowel + Vowel:Language	3034	9.94	1	< 0.002
4b Token^3 ∗ Token ∗ Vowel + Vowel:Language	3031	22.2	6	< 0.002
5 Random: (1+Vowel | Subject)	3067	62.8	15	< 0.001

for the effect of rounding. The pattern is in line with previous findings (Mann & Repp, 1980; Mitterer, 2006; Smits, 2001).

Table 3 shows AIC values and the results of likelihood ratio tests that went into selecting the mixed effect model for this experiment's data. A relatively maximal random effects structure (TOKEN ∗ VOWEL | SUBJECT) was used with models 1–5 to explore the experimental fixed effects. Starting with a baseline model that includes only TOKEN number (centered and treated as a continuous variable), we found that adding VOWEL predictor variables improved the fit substantially. The fit was further improved by adding a fixed effect for the LANGUAGE of the listener, though as lines 3a and 3b show a model without the LANGUAGE main effect (3b) gave a slightly better fit than did a model with a main effect and interaction. (The statistics for models 3a and 3b in the table are for comparisons against model 2.) Including a cubic term for TOKEN improved the fit substantially (model 4a), with a slightly better AIC when allowed to interact with TOKEN and VOWEL. Finally, with model 5 we found that simplifying the random effect structure (removing by-subject random slopes for TOKEN) significantly reduced the accuracy of the fit. So, we are reporting here the structure of model 4b, the best fitting logistic mixed-effects model that we were able to fit. (We were not able to fit full models with TOKEN entered as a categorical variable, because of the quite substantial increase in the number of model coefficients that is required in these models.)

Table 4 shows the random effects structure and the fixed effect coefficients in the best model (#4b in Table 3 – Token^3 ∗ TOKEN ∗ VOWEL + VOWEL:LANGUAGE) of the experiment 1 results. Negative coefficients are associated with greater probability of 'sh'-responses and positive coefficients with greater probability of 's'-response. As shown, there was a significant negative effect of TOKEN number, suggesting that the RESPONSE is more likely to be 'sh' as Token number increases. The round vowel [u] had a significant positive coefficient reflecting the patterns seen in Fig. 5 that responses are more likely to be 's' before [u] than before [a] or [y].

Although we did not find a significant effect of LANGUAGE itself, there was a significant effect of the VOWEL:LANGUAGE interaction which is apparent in Fig. 5 as the greater difference in the different identification for vowels [a], [u], and [y] in French than in English. The [u] by French coefficient is reliably different from zero indicating an increase in the number of 's' responses for French speaking listeners, and the [y] by French coefficient is marginally reliable ($p = 0.064$), suggesting the French-speaking listeners' response in the [y] environment was different from English-speaking listeners.

We analyzed the identification boundaries (see Fig. 6) in a repeated-measures analysis of variance and this analysis found a pattern that is compatible with the results found in the mixed-effect logistic regression. Note that with a single source of random effect (subjects) the repeated-measures analysis of variance is equivalent to the maximal mixed-effects model (Barr et al., 2013). There was a main effect for VOWEL ($F[2,80] = 32.1$, $p < 0.001$) and a VOWEL by LANGUAGE interaction ($F[2,80] = 2.49$, $p < 0.02$). Planned comparisons of the boundary locations found that for French-speaking listeners both [u] and [y] boundaries were reliably different from the location of the [a] boundary ([u] vs. [a]: $t[20] = 7.6$, $p < 0.001$; [y] vs. [a]: $t[20] = 3.1$, $p < 0.01$) while for English-speaking listeners the boundary for [u] was different from

Table 4
Experiment 1: the random effects structure and fixed effects coefficients of the final mixed effects logistic model. TOKEN, VOWEL, and the TOKEN:VOWEL and VOWEL:LANGUAGE interactions are included as fixed-effects terms. Number of observations: 7441, groups: Subject, 42.

Random effects:

Name	Variance	Std. dev.	Corr				
(Intercept)	1.95	1.40					
Token	0.16	0.40	−0.57				
Vowely	0.98	0.99	0.17	−0.03			
Vowelu	0.52	0.72	0.25	−0.53	0.53		
Token:Vowely	0.04	0.20	0.31	0.41	0.37	0.25	
Token:Vowelu	0.20	0.45	−0.33	−0.26	−0.50	0.38	−0.24

Fixed effects:

Name	Estimate	Std. error	z value	Pr($< \|z\|$)
(Intercept)	−0.747	0.310	−2.413	0.016 *
Token^3	0.010	0.015	0.740	0.459
Token	−1.925	0.131	−14.700	0.001 ***
Vowely	0.402	0.277	1.453	0.146
Vowelu	1.233	0.231	5.349	0.001 ***
Token^3:Token	0.008	0.003	2.945	0.003 **
Token^3:Vowely	0.016	0.019	0.840	0.401
Token^{3}Vowelu	0.027	0.018	1.493	0.135
Token:Vowely	−0.215	0.162	−1.324	0.186
Token:Vowelu	−0.250	0.164	−1.523	0.128
Vowela:langFrench	0.391	0.399	0.981	0.326
Vowely:langFrench	0.909	0.454	1.850	0.064
Vowelu:langFrench	1.371	0.454	3.019	0.003 **
Token^3:Token:Vowely	−0.004	0.003	−1.287	0.198
Token^3:Token:Vowelu	−0.009	0.003	−2.785	0.005 **

*p < 0.05; **p < 0.01; ***p < 0.001.

[a] ($t[20] = 4.46$, $p < 0.001$), but the boundary for [y] was not ($t[20] = 1.35$, $p = 0.096$). Incidentally, for the English-speaking listeners the [u] and [y] boundaries were reliably different ([u] vs. [y]: $t[20] = 2.33$, $p < 0.05$), which was also the case for the French-speaking listeners ([u] vs. [y]: $t[20] = 3.98$, $p < 0.001$).

2.3. Discussion

Experiment 1 found that both English-speaking and French-speaking listeners show perceptual compensation for rounding coarticulation. However, the results indicate that English-speaking listeners were less sensitive to the effects of lip rounding, with a smaller compensation effect for [u] and no detectable compensation effect for [y]. The nominal boundary locations followed the same order for both groups: [u] > [y] > [a]; however, the significant interaction of LANGUAGE and VOWEL suggests that the size of the rounding effect was different for French and English listeners. The statistical analysis showed that the English-speaking listeners' pattern of boundary locations was [u] > [y],[a], while the French-speaking listeners' pattern was [u] > [y] > [a].

It is interesting that the compensation effects follow the same nominal order for both listener groups. This may reflect a language-independent auditory contrast effect in the perception of fricatives. The idea is that the lowered formant values of the vowel in [u] contrast with the fricative spectrum and increases the perceived center of spectral energy, resulting in a greater number of 's' responses. This account is problematic on a couple of counts. First, backward masking effects are generally seen as relatively weak, so if this compensation effect is to be seen as a masking effect it must be more like an informational masking effect. Second, [y] is acoustically a lot more like [i] than like [u]. The major difference between [i] and [y] is in the F2/F3 ratio, mostly F3. Thus, the compensation effect with [y] is very hard for a simple auditory contrast model to explain.

Our conclusion is that both phonetic and phonological factors may be involved in this pattern of results. Phonologically, because French has both [u] and [y] as native phonemes, the French-speaking listeners in this experiment were perceptually sensitive to the rounding of both [u] and [y], while for English-speaking listeners who have [u] in their native phonology but not [y], the lip rounding of [y] was not as salient or coherent for these listeners, and thus less likely to evoke the perceptual compensation effect. Phonetically, because the back round vowel [u] in English involves a smaller degree of lip-rounding than is found in French or German, English-speaking listeners may have expected a smaller amount of labial coarticulation between the consonant and vowel than did the French-speaking listeners, and thus produced a smaller compensation effect for [u].

One thing that is quite clear from these results is that the linguistic experience of the listener had an impact

on perceptual compensation for coarticulation. What is more, the perceptual pattern of responses reflects the phonetic and phonological realities of the listener's native language. English-speaking listeners' responses are in line with the limited phonetic and phonological status of vowel lip rounding in English, in contrast with the larger compensation effect observed with French-speaking listeners which is in line with the more important status of rounding in French [u] and [y].

3. Experiment 2

Experiment 2 is a test of how malleable compensation for coarticulation is for English-speaking listeners. The experiment tested whether English-speaking listeners can use the enhanced degree of lip rounding in [u] and [y] that was actually involved in the production of the German stimuli that we used in experiment 1 when they can see the speaker round her lips during the vowel (and perhaps see the anticipatory lip coarticulation as well). The question is whether English-speaking listeners will show a pattern more like French-speaking listeners when they can see that [u] has very rounded lips, and that the vowel [y] (which has formant frequencies close to those in [i]) also has very rounded lips. Traunmüller and Öhrström (2007) found that lip rounding is strongly signaled in visual displays. Also, Mitterer (2006) used visual lip rounding to induce a vowel rounding percept which produced perceptual compensation in fricative identification. If, as Experiment 1 suggested, the compensation effect is dependent upon the native-language experience of the listener, then the strongest test of the language-specific basis of compensation is to present both audio and visual lip rounding. If English-speaking listeners continue to be relatively insensitive to vowel rounding, and show no compensation effect in the [y] environment, then we would have to conclude that compensation for coarticulation is strongly mediated by linguistic experience even when clear gestural information is available. On the other hand, we may find that the compensation response is much stronger when vowel rounding information is given in the audio/visual speech signal. This would indicate that listeners have the ability to use visual lip gestures in perceptual compensation for coarticulation regardless of linguistic experience with lip rounding in the native language. Thus, the basic question addressed by Experiment 2 is whether the phonetic mode of listening is gestural or whether it is more linguistic. Ettlinger and Johnson (2009) framed this question as one of feature perception versus exemplar-based perception.

3.1. Methods

3.1.1. Subjects

Thirty-nine listeners between ages 19 and 27 without any speech or hearing problem participated in the audio-visual experiment. All participants were native speakers of American English who were attending the University of California, Berkeley at the time of participation. Several participants were bilinguals or equally fluent in other languages including Hindi, Spanish, and Farsi, but none of them spoke a language with front round vowels.

3.1.2. Stimuli

In order to see the effect of visual lip rounding in compensation for coarticulation, the auditory stimuli from Experiment 1 were each used as the sound track in the original videos of the face of the model speaker articulating [s]V and [ʃ]V syllables. The talker who recorded these stimuli repeated each token three times, and we selected tokens for Experiments 1 and 2 with natural face movement and audio which were comparable in duration and loudness. One native French speaker confirmed that each video clip looked natural and can be used as the representative as the production of these CV syllables in French. The vowel portions in the audio and video stimuli always matched: the audio of vowel [a] was aligned with the face articulating [a], etc. Thus, the audio tokens with [a] were played with a face that showed relative unrounded lips during the vowel, while the tokens with [u]/[y] were played with movies that had rounded lips during the visual vowel.

In order to test the effects of visual fricative cues, three different fricative movies were used for each vowel environment. The original movies of the face saying [s]V and [ʃ]V were aligned at the CV transition to be synchronous with the corresponding audio stimuli in the [s]/[ʃ] continuum used in experiment 1. A third movie for each vowel environment was made by blending the [s]V and [ʃ]V movies using the dynamic morphing function in WAX (Satish, 2012). For each vowel environment, we produced a blended movie that interpolated frame by frame between the movies for [s] and [ʃ]. In order to do this we outlined the lips in each frame of each movie so that the interpolation was anchored on the lips. Thus, for example, the visual stimuli for the vowel [a], were movies of [sa], [ʃa], and the [sa]/[ʃa] blend.

The visual difference between [s] and [ʃ] was not very conspicuous before [u] and [y] but the fricatives are noticeably different before the unround vowel [a] where [ʃ] has rounded lips and [s] has unrounded lips. Before the round vowels [u] and [y] this fricative difference was largely eliminated by lip rounding coarticulation from the vowel. The impact of this will be addressed in more detail in the discussion section.

3.1.3. Procedure

The participants saw the stimulus movies on an LCD monitor at a distance of about 20 in. and heard the audio portion over headphones (AKG K240) at a comfortable listening level. The subjects were asked to identify the initial consonant as either 's' or 'sh' by pressing a keys on a standard computer keyboard. To shorten the duration of the experiment, the two endpoints from the nine-step continua (Token 1 and Token 9) were removed from the list. As a result, the participants responded to 441 visual tokens (7 audio fricatives × 3 vowels × 3 visual fricatives × 7 repetitions). The list of tokens was randomized separately for each participant. To divide the experiment into two blocks, the participants were invited to take a short break after trial number 220.

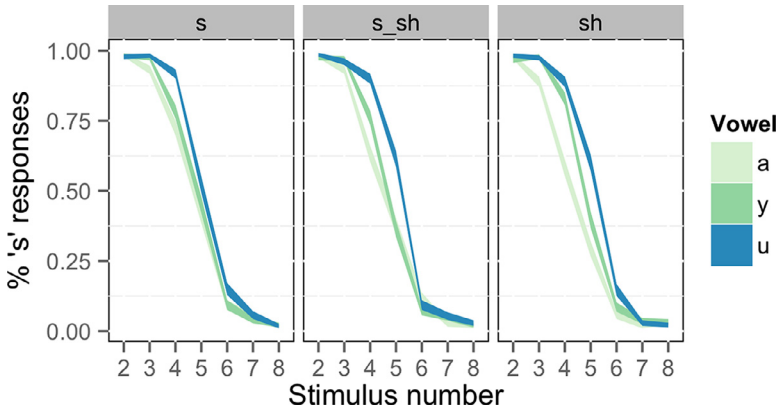

Fig. 7. Experiment 2 results. Identification curves showing proportion of 's' responses averaged across listeners as a function of the stimulus vowel for each of the visual fricative conditions. The width of each ribbon corresponds to the standard error between subjects.

3.1.4. Statistical analysis

The analysis method used for Experiment 1 was also adopted for this experiment. In the logistic mixed-effects regression, the dependent measure was the subject's RESPONSE on each trial and the independent predictors were (1) the fricative TOKEN number, centered and the cube of the token number, (2) the context VOWEL ([a], [u], or [y]), and (3) the VISUAL FRICATIVE ([s], [s/ʃ], [ʃ]).

As with the data of Experiment 1, we calculated the 50% crossover boundaries by fitting a four-term logistic equation to the probability 's' identification functions. These data were analyzed using a repeated measures analysis of variance with the same predictor variables used in the logistic mixed-effects regression.

Table 5
Experiment 2 model selection. The χ^2, degrees of freedom, and p-Values in this table are for likelihood ratio tests as discussed in the text.

Model	AIC	χ^2	df	p-Value
1 Token^3 + Token	9063	–	–	–
2 Token^3 + Token ∗ Vowel	8512	595	22	< 0.001
3 Token^3+ Token ∗ Vowel + VideoFricative	8502	13.6	2	< 0.002
4 Token^3+ Token ∗ Vowel ∗ VideoFricative	8493	28.8	10	< 0.002
Random: (1 + Token ∗ Vowel ∗ Video | Subject)	8696	97	150	= 0.999
Random: (1 + Token ∗ Vowel | Subject) + (Video | Subject)	8500	4.9	6	< 0.552
Random: (1+Vowel | Subject)	8812	349	15	< 0.001
4a Token ∗ Vowel ∗ VideoFricative	8533	41.2	1	< 0.001

3.2. Results

The mean proportion of 's' responses across all subjects for each acoustic fricative token along the continuum in the three different vowel environments for each visual fricative condition is plotted in Fig. 7. As in Experiment 1, the unround vowel /a/ yielded fewer 's' responses than the rounded vowel /u/. The number of 's' responses for vowel /y/ appears to be slightly more than that for vowel /a/ but less than that of vowel /u/. In short, as in Experiment 1, the results seem to show a greater compensation effect with the vowel /u/ with a weaker effect for the vowel /y/.

The best fitting model included main effects and interactions for TOKEN, VOWEL, and VIDEO FRICATIVE. (model 4 in Table 5). As in the analysis of experiment 1, we found that models that included a nonlinear token effect fit the data better (comparing model 4 with model 4a [$\chi^2(1) = 41.2$, $p < 0.001$]). The best fitting random effects structure for these data had random intercepts for each subject and random slopes for the VOWEL by TOKEN interaction. There was no improvement by adding random slopes for VIDEO FRICATIVE (the maximum model), but the fit was substantially improved by having slopes for the TOKEN by VOWEL interaction as opposed to just having random slopes for the different vowels. The estimated values for the random effects and fixed-effect predictors in the full model are listed in Table 6.

As in Experiment 1, there was a significant effect of TOKEN: the negative TOKEN coefficient indicates that responses were more likely to be "sh" as token number increased. Also paralleling experiment 1, these data show a reliable vowel effect for [u] but not for [y], which is reflected in the TOKEN by VOWEL interaction as well. The effect of the VISUAL FRICATIVE was to increase the number of "sh" responses (negative coefficient) with the [ʃ] and [s/ʃ] videos. The only coefficient in the model that showed an effect for [y], differentiating it from the default vowel [a], was in the VOWEL by VIDEO interaction. We will argue in the discussion that this is an effect of visual fricative rounding and not an effect of compensation for coarticulation of the vowel rounding on the fricative identification.

The category boundaries were estimated as in Experiment 1 for each subject in each condition (9 boundaries per subject - three levels of VOWEL crossed with three levels of VISUAL FRICATIVE). Fig. 8 shows the distributions of these boundary estimates. The results of a repeated-measures analysis of variance of these boundary data are consistent with the logistic regression analysis. There was

Table 6
Mean estimates of mixed effects logistic model of Experiment 2. Token, Vowel, Visual Fricative (VF), and Vowel:VF interaction were included as fixed-effects terms. Number of obs: 17,640, groups: Subject, 39.

Random effects:							
Name	Variance	Std. dev.	Corr				
(Intercept)	2.34	1.53					
Token	0.31	0.56	0.27				
Vowely	1.15	1.07	−0.59	−0.06			
Vowelu	1.05	1.03	−0.45	−0.35	0.61		
Token:Vowely	0.09	0.30	0.00	−0.13	0.32	0.32	
Token:Vowelu	0.36	0.60	−0.26	−0.08	0.18	0.39	0.64

Fixed effects:

Name	Estimate	Std. error	*z* value	Pr(<|z|)
(Intercept)	−0.684	0.262	−2.612	0.009 **
Token^3	0.053	0.008	7.087	0.001 ***
Token	−2.238	0.122	−18.371	0.001 ***
Vowel.y	0.194	0.213	0.909	0.363
Vowel.u	0.925	0.211	4.389	0.001 ***
Video.s/ʃ	−0.241	0.119	−2.016	0.043 *
Video.ʃ	−0.676	0.125	−5.428	0.001 ***
Token:Vowel.y	−0.137	0.115	−1.197	0.231
Token:Vowel.u	−0.410	0.150	−2.726	0.006 **
Token:Video.s/ʃ	0.033	0.088	0.380	0.704
Token:Video.ʃ	0.018	0.089	0.197	0.844
Vowel.y:Video.s/ʃ	−0.011	0.170	−0.063	0.950
Vowel.u:Video.s/ʃ	0.264	0.175	1.508	0.132
Vowel.y:Video.ʃ	0.641	0.174	3.693	0.001 ***
Vowel.u:Video.ʃ	0.722	0.179	4.034	0.001 ***
Token:Vowel.y:Video.s/ʃ	0.001	0.130	0.008	0.993
Token:Vowel.u:Video.s/ʃ	−0.025	0.138	−0.178	0.859
Token:Vowel.y:Video.ʃ	−0.011	0.132	−0.083	0.934
Token:Vowel.u:Video.ʃ	−0.051	0.140	−0.364	0.716

*p < 0.05; **p < 0.01; ***p < 0.001.

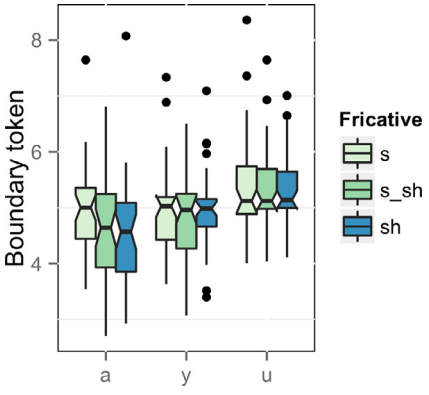

Fig. 8. Experiment 2 results. The distributions of the identification function 50% boundaries as a function of the visual fricative for each of the context vowels.

Table 7
Pairwise comparisons of 50% cross-over boundaries as a function of the visual fricative context for different vowel comparisons.

		t value	df	*p*-Value
s video	a - u	3.98	38	<0.001**
	a - y	0.45	38	0.328
	u - y	3.19	38	<0.01*
s_ʃ video	a - u	5.56	38	<0.001**
	a - y	1.62	38	0.057
	u - y	5.04	38	<0.001**
ʃ video	a - u	6.36	38	<0.001**
	a - y	3.52	38	<0.001**
	u - y	3.54	38	<0.001**

a VOWEL main effect ($F[2,76] = 27.87$, $p < 0.001$), and a VIDEO main effect ($F[2,76] = 7.57$, $p < 0.01$), and the VOWEL by VIDEO interaction was marginally significant ($F[4,152] = 2.49$, $p = 0.0454$).

Responses to [u] were different from [a] in all three fricative movies, while responses to [y] were different from [a] in only the [ʃ] movie (Table 7). Thus, with [y] it appears that the rounding of the fricative [ʃ] had an impact on perception but the coarticulatory rounding of the vocalic [y] portion of the stimulus did not. Fig. 8 shows that the main consequence of seeing the face in these AV stimuli was that coarticulatory differences on the vowel [a] led to a different boundary on the [s]-[ʃ] continuum. The main visual difference between [s] and [ʃ] was in the [a] context and this is where we see the largest effect of the visual fricative. The vowel compensation effect was virtually unchanged by visual lip rounding.

3.3. Discussion

The result of Experiment 2 is that English-speaking listeners compensated for [u] reliably, but that they still didn't compensate for the rounding of [y] even with the availability of visual lip rounding information. This result suggests that compensation is driven by listeners' phonetic expectations which are not over-ridden even when phonetic reality (the presence of visual vowel lip rounding together with an acoustically backer vowel) is inconsistent with those phonetic expectations.

The visual fricative effect in this experiment is important because it demonstrates that listeners were attending to and using visual phonetic information as they performed the fricative identification task. Experiment 2 found that American English-speaking listeners are sensitive to visual lip rounding information when deciding that a fricative is [s] or [ʃ]. The boundary for the [sa]-[ʃa] continuum shifted depended on the identity of the visual fricative. We also found that compensation for the vowel context was present for [u] regardless of the visual fricative. The listeners' response to the visual displays of [y] were different from [u], showing no effect of vowel rounding on the identification of the preceding fricative.

To further explore the effect of visual vowel rounding we statistically compared the results of experiment 1 and 2. The calculated identification boundaries of English-speaking listeners from Experiment 1 (audio-only modality) and Experiment 2 (audiovisual modality) were combined and analyzed together in a repeated-measures analysis of variance. The predictor variables were MODALITY(Audio vs. Audiovisual) and VOWEL ([a], [y], or [u]). For this analysis we used only responses to the blend fricative ('s_sh') because we were interested in the effect of vowel lip rounding on fricative perception - the compensation effect. The same result was obtained with the other two visual fricative conditions as well. There was a VOWEL main effect ($F[2,116] = 25.9$, $p < 0.001$). Post-hoc comparisons of the vowel effect differences have been reported earlier with each experiment. The MODALITY main effect was not significant, nor was the MODALITY by VOWEL interaction (both Fs < 1). This comparison confirms the impression that the compensation for compensation effect was unchanged by the addition of visual information in experiment 2.

4. General discussion

4.1. Language-specific compensation for coarticulation

We found that participants were more likely to respond 's' in front of rounded vowels and more likely to respond 'sh' in front of the unround vowel [a], hence replicating the effect of compensation for coarticulation that has been reported previously. However, not all rounded vowels produced this effect equally. The vowel [u] consistently yielded more 's'-responses across all listeners regardless of their native language background. This may be partly attributable to the lower F2 of [u] versus the raised F2 of [y] in which the effect of rounding (lowering of formants etc.) is more salient than in [y]. According to the spectral contrast account of compensation for coarticulation, the contrast in the formants between fricative and vowel can cause listeners to register a different fricative depending on the spectrum of the neighboring vowel. However, spectral contrast is a poor explanation for the result for [sy]. Although [y] is spectrally similar to [i], the boundary shift for [y] is similar to the pattern of [u] by both listener groups; i.e., [sy] behaves like [su] for the purposes of rounding compensation, but is acoustically similar to [si]. The major acoustic difference between [i] and [y] is in F2/F3 frequency, but both contain ample high-frequency energy, whereas [u] does not. An [u]-like compensation effect with [y] is difficult for auditory contrast to explain given its acoustic similarity to [i]. A possible alternative under the spectral contrast approach would be to see the effect in terms of the size of contrast. Since the acoustic contrast between the fricative and vowel spectrum is generally larger for [u] than it is for [y], one would expect a weaker effect for [y] than for [u] in rounding compensation as we find in the experiment. Yet Viswanathan et al. (2013) using tone-speech contexts show that observed effects were sometimes greater for smaller frequency difference between a precursor and target, which makes the account not intuitively plausible.

Another difficulty for any explanation of compensation for coarticulation that sees it rooted in universal perception processes (whether auditory or gesture recovery) is the fact that linguistic experience modulated compensation in this study. In Experiment 1, we found that the French-speaking listeners had stronger compensation responses for [u] and [y] than did English-speaking listeners. Since /y/ is not a native phoneme of English, English listeners are familiar only with [u], whereas French listeners whose native language contains /y/ in its phoneme inventory are familiar with both [u] and [y]. The French listeners' greater compensation effect might be attributable to the role of rounding in the French vocalic inventory. As [u] and [y] are contrastive, French listeners must rely on rounding to distinguish [y] from [i], which shares other characteristics with [y] derived from place and height. On the other hand, rounding is redundant for back vowels in English. If we discuss the contrast in terms of a phonological feature, the round vowels in English instead can be described with only place and height features. The different phonological status of rounding in French and English might have led the listeners to have differential sensitivity toward rounding, which in turn eventually resulted in the differential degree of compensation effect to rounding. The result is taken as evidence that familiarity to phonemes, or possibly 'features', that are contrastive in the native language can affect listeners' ability to compensate. Of course, the study by Ettlinger and Johnson (2009) reports that knowing a featural contrast (tense/lax) does not necessarily extend to an unfamiliar sound, implying that having experience with a phonetic segment like [u], is more important than experience with a phonetic feature like [round].

The results of Experiment 2 further support the conclusion that compensation for coarticulation is language-specific.

Although the participants could see the speaker rounding her lips as she produced [u] and [y], the number of 's'-responses by English listeners did not increase significantly for either vowel. If the compensation effect depends on the perception of a lip rounding gesture as suggested by a gesture recovery account, compensation in the [y] context should have increased with addition of visual information. The result suggests that language-specific knowledge cannot be easily supplemented by seeing how the unfamiliar sound is articulated.

4.2. Role of visual information in compensation for coarticulation

We tested whether compensation for coarticulatory lip rounding extends to unfamiliar sounds when visual information is presented to English listeners. We did not find any increase of compensation with videos of either [u] or [y] in English. Mitterer (2006) apparently had a different result. He found that Dutch listeners compensated more for [y] in audiovisual presentation as compared with audio-only stimuli.It is important to note though that the acoustic vowel signals in Mitterer's study were ambiguous while our acoustic vowels were clearly [u] and [y]. If the listeners' weak effect and greater variability for compensation for [y] in Experiment 1 is attributable to their incapability to ascertain the lip rounding in the unfamiliar vowel, we predicted that they would show compensation for [y] when they could see the speaker's rounded lips. Aside from the non-native vowel [y], it was equally plausible to predict that compensation with [u] would increase as well with visual presentation because the German [u] appears to have more lip rounding than English [u]. However, compensation for neither [u] nor [y] was enhanced as a result of adding visual information.

One remote possibility is that the participants in Experiment 2 managed to complete the task without referring to visual information. i.e. listeners might not have given sufficient attention to the visual stimulus. However, we know that the participants were attending to the visual stimuli because there was a significant effect of the visual fricative, especially with vowel [a] where visual [s] was most different from visual [ʃ]. The participants were more likely to respond 's' when they saw the [s] visual fricative and more likely to respond 'sh' when they saw /s_ʃ/ (blend) visual fricative, and even more so for /ʃ/ visual fricative. This effect of the visual fricative was not present in the [u] and [y] vowel conditions where anticipatory vowel lip rounding reduced the visual difference between /s/ and /ʃ/. The visual difference between [s] and [ʃ] which is quite pronounced before [a] is greatly reduced when they are followed by a rounded vowel. The results thus suggest that the participants were attending to, and using the visual input in this experiment. The results also imply that integration of visual and audio information during speech perception occurs even when the acoustic signal is relatively clear; while simultaneously, robust visual information may fail to make an impact when the speech sound that is being presented is unfamiliar. [We were tempted to say that the integration of audio and visual information is automatic, but a reviewer pointed us to an important study that indicates that this is not the case (Alsius et al., 2014).] Therefore, lack of a visual effect on compensation for both [y] and [u] supports the possibility that compensation is also driven by knowledge of the native language, and is not simply modulated by veridical gestural perception. It may be that perception of visual information is also affected by linguistic experience. i.e. although a listener may 'see' lip rounding, this feature may not register in the percept because it doesn't contribute to a familiar sound. This may also explain why we did not see an increase of the compensation pattern for the native vowel [u]; English-speaking listeners may have already compensated for lip rounding in the audio-only stimuli as much as they were going to, so the visual vowel added no new essential information.

4.3. Where does direct realism stand with respect to these findings?

In the introduction we described three factors that are likely to be involved in speech perception: properties of auditory transduction, phonetic knowledge of speech production, and lexical knowledge of phonological patterns. In connection with the second of these, phonetic knowledge, we cited Liberman and Mattingly (1985) who held that phonetic knowledge is innate, and Best (1995) who assumed, we think rightly, that phonetic knowledge is learned. We also suggested that the version of direct realism associated primarily with Fowler (1986), Fowler (1996) and Goldstein and Fowler (2003) is more in line with Liberman and Mattingly's view of an innate speech perception capability. The results of our experiments are compatible with the Best (1995) view – that phonetic knowledge is learned. Although this aspect of direct realism is not central to our concerns in this paper, it should be addressed in the interest of correctly ascertaining the implications of our results for speech perception theory.

Gibson's (1966) statement of direct realism envisioned the senses as perceptual systems that change in response to the environment. He said, "A perceiver is a self-tuning system. What makes it resonate to the interesting broadcasts that are available instead of to all the trash that fills the air? The answer might be that the pickup of information is reinforcing. [...] A system 'hunts' until it achieves clarity." (p. 271).

Perceptual learning is not emphasized in Fowler's various presentations of a direct realist theory of speech perception (Fowler, 1986; 1996; Fowler and Housum, 1987; Galantucci et al., 2006; Goldstein and Fowler, 2003). Indeed, the emphasis in her direct realism is on the richness of the phonetic signal not on the "trash".

For example, Fowler (1986, p. 15) discussed the topic of perceptual learning in the context of top-down influences on perception saying,

> "It is not that an event theory of speech perception has nothing to say about perceptual learning. [...] However, what is said is not yet well enough worked out to specify how, for example, lexical knowledge can be brought to

bear on speech input from a direct-realist, event perspective." (Fowler, 1986, p. 15).

In further elaboration on this point she said,

"I prefer a similar approach [...] that makes a distinction between what perceivers can do and what they may do in particular settings. [...] there is a need for the informational support for activity to be able to be directly extracted from an informational medium and for perception to be nothing other than direct extraction of information from proximal stimulation." Fowler (1986, p. 15).

These comments about perceptual learning suggest a view of perception in which perceivers learn what to do with perceptual results, but perception itself is direct, specified by the world and not elaborated in any way by knowledge obtained through experience.

This sense of perception as "of" the world is heightened by Fowler and Dekle (1991) discussion of their finding that listeners are able to use haptic sensation in speech perception despite a lack of experience with this modality – i.e. that experience is not needed when perception is direct. This "universality" of speech perception is further highlighted in comments on the evolutionary basis of perception:

"These perceptual systems were shaped by *natural selection* to serve the function of acquainting perceivers with components of their niches. Auditory perception can only have been selected for the same function. There is no survival advantage to hearing structured air, but there is an advantage, for example, to locating a large lumbering animal out of view and to detecting which way, in respect to one's self, it is lumbering." (Fowler, 1986, p. 1732).

For Fowler, then, evolution drives development of a universal perceptual system that links speech gestures in the mouth with speech gestures in the ear. This introduces the notion of parity – that speech perception and production trade in the same material, speech gestures. Indeed, Fowler (1996) says as much:

"In the theory, listeners perceive gestures because perceptual systems have the function universally of perceiving real world causes of structure in media, such as light, air, and the surfaces of the body, that sense organs transduce. Accordingly, perception is generally heteromorphic with respect to structure in those media; instead, perception is *not just homomorphic* with, it is of, the real-world events that cause the structure. That is, speech perceivers, and perceivers in general, are realists (Fowler and Housum, 1987). Indeed, it is their status as perceptual realists that explains parity." (Fowler, 1996, p. 1731)

Perhaps it makes sense then that readers encountering these descriptions of the direct-realist theory of speech perception would think of it as a universal theory of perception, in which humans have evolved to have an innate ability to perceive the gestures of speech. This is clearly not what is intended by at least some proponents of direct realism. Best (1995) calls this "the misconception that the direct realist view of speech perception is incompatible with perceptual learning" (p. 180).

Goldstein and Fowler (2003, p. 26) attempted to explain the role of perceptual learning in direct realism:

"Readers unfamiliar with direct realism sometimes ask how infants learn to connect the patterns of stimulation at their sense organs with the properties that stimulation specifies in the world. This question, however, has to be ill-posed. The only way that perceivers can know the world is via the information in stimulation. Therefore, the infant cannot be supposed to have two things that need to be connected: properties in the world and stimulation at sense organs. They can only have stimulation at the sense organs that, from the start, gives them properties in the world. This is not to deny that perceptual learning occurs (e.g., Gibson and Pick, 2000; Reed, 1996). It is only to remark that learning to link properties of the world with stimulation at the sense organs is *impossible* if infants begin life *with no such link* enabled."

We understand this to mean that learning in direct realism is "tuning" a system that in its basic function is an innately specified result of evolutionary development. In light of results like those reported by Lewkowicz et al. (2015) on the slow emergence of audio-visual coherence in speech perception by very young children, one wonders whether the perceptual link between audition and gestures is really as innate as Goldstein and Fowler envision. Their view of the innateness of a link between properties of the world with stimulation of the sense organs is a much more general theory of innateness than the concept of a speech module that Liberman and Mattingly (1985) advocated. In this section though we just want to remark on how little attention is devoted in Fowler's variant of direct-realism to questions of perceptual learning.

It is no wonder, then, that reviewers such as Diehl et al. (2004) can choose to describe an alternative to direct-realism as a "General auditory and *learning* approach" to speech perception, as if learning was not a feature of direct realism. In fact, the innate link hypothesis, proposed by Fowler (1996) and Goldstein and Fowler (2003) is an extreme position where direct-realism is most distinctly different from cognitive/neural theories of perception (see Fowler, 1996, p. 1732 on "public" and "covert" aspects of perception).

The version of direct-realism described in Best (1995) is less noticeably different from cognitive/neural theories in this regard. According to Best (1995, p. 180), learning is: "experience-based attunement to detecting higher-order invariants of objects, surfaces, and events" which "increases economy in information pick-up", and "increases specificity and differentiation of the critically distinctive information characterizing one object or event as different from another". So on this view, the effect of experience forms a basis for perception of acoustic properties of events and objects and to (1) speed the perception of familiar speech sounds, and (2) increase accuracy in the perception of familiar speech sounds.

One could argue that the English-speaking listeners' failure to show a compensation effect for [y] in our study (even

with the availability of visual lip rounding information in Experiment 2) reflects a lack of "experience-based attunement to detecting higher-order invariants" in the speech signal. In this regard our results are completely compatible with Best's (1995) version of direct realism. It is also interesting that familiar gestures in unfamiliar configurations were problematic for our listeners. Lip rounding and tongue fronting are familiar to English listeners, but [y] is not. The finding is also quite compatible with Best's (1995) conception of direct realism because for her the objects of perception are language-specific gestural constellations rather than "simple gestures" (p. 189).

5. Conclusion

We have shown that compensation for coarticulation is language-specific and visual perception of speech does not itself change listeners' pattern of compensation. The findings suggest that compensation for coarticulation is a phenomenon that is modulated not only by sensory factors like spectral contrasts between segments or the phonetic interpretation of the specific gestures associated with segments, but also by phonetic knowledge of one's native language. Additionally, our result on audiovisual modality raises an important question about the perception of speech gestures: If knowing articulatory gestures is directly linked to the phonological knowledge of sounds, the perception of the gesture may be language-specific such that it is only applicable to the sounds within one's native language. Our result does not offer a conclusive answer to this question, and further research is needed to develop an understanding of how and which visual properties are used by listeners during compensation. Finally, our cross-linguistic comparisons shed light on the role of linguistic experience shapes spoken language processing.

Acknowledgments

We are grateful to two anonymous reviewers, Holger Mitterer, and the members of Phonology Lab at UC Berkeley Linguistics Department for their valuable comments on this paper. We presented an earlier version of this project at the 164th meeting of the Acoustical Society of America and we thank participants there for their feedback. Many thanks to Elsa Spinelli for help in data collection in France, and to Nicholas Strzelczyk and Carson Miller Rigoli for their work in helping us cr eate the stimuli and collect perception data in Berkeley. This work was funded by the National Science Foundation (BCS1147583) and the National Institute of Deafness and Communication Disorders (R01DC011295-3).

References

Alsius, A., Möttönen, R., Sams, M.E., Soto-Faraco, S., Tiippana, K., 2014. Effect of attentional load on audiovisual speech perception: evidence from ERPS. Front Psychol. 15 July 2014 http://dx.doi.org/10.3389/fpsyg.2014.00727

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing; keep it maximal. J. Mem. Lang. 68, 255–278.

Beddor, P.S., Harnsberger, J.D., Lindemann, S., 2002. Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. J Phon. 30, 591–627.

Beddor, P.S., Krakow, R.A., 1999. Perception of coarticulatory nasalization by speakers of english and thai: evidence for partial compensation. J. Acoust. Soc. Am. 106, 2868–2887.

Benguerel, A.P., Adelman, S., 1976. Perception of coarticulated lip rounding. Phonetica 33, 113–126.

Benguerel, A.P., Cowan, H.A., 1974. Coarticulation of upper lip protrusion in french. Phonetica 30, 41–55.

Best, C.T., McRoberts, G.W., Sithole, N.M., 1988. Examination of perceptual reorganization for nonnative speech contrasts: zulu click discrimination by english-speaking adults and infants. J. Exp. Psychol. Hum. Percept. Perform. 4, 45–60.

Best, C.T., 1995. A direct realist view of cross-language speech perception. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. York Press, Timonium, MD, pp. 171–204.

Bond, Z.S., 2005. Slips of the ear. In: Pisoni, D.B., Remez, R.E. (Eds.), The Handbook of Speech Perception. Blackwell Publishing, Malden, MA, pp. 290–310.

Bradlow, A.R., Bent, T., 2002. The clear speech effect for non-native listeners. J. Acoust. Soc. Am. 112 (1), 272–284.

Bregman, A.S., 1990. Auditory scene analysis: the perceptual organization of sound. The MIT Press, Cambridge, MA, pp. 1–792.

Byrd, D., Saltzman, E., 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. J. Phonetics 31, 149–180.

Diehl, R.L., Lotto, A.J., Holt, L.L., 2004. Speech perception. Annu. Rev. Psychol. 55, 149–179.

Diehl, R.L., Walsh, M.A., 1989. An auditory basis for the stimulus-length effect in the perception of stops and glides. J. Acoust. Soc. Am. 85, 2154–2164.

Elman, J.L., McClelland, J.L., 1988. Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. J. Mem. Lang. 27 (2), 143–165.

Ettlinger, M., Johnson, K., 2009. Vowel discrimination by english, french and turkish speakers: evidence for an exemplar-based approach to speech perception. Phonetica 66, 222–242.

Forrest, K., Weismer, G., Milenkovic, P., Dougall, R.N., 1988. Statistical analysis of word-initial voiceless obstruents: Preliminary data. J. Acoust. Soc. Am. 84, 115–123.

Fowler, C.A., 1986. An event approach to the study of speech perception from a direct-realist perspective. J. Phon. 14, 3–28.

Fowler, C.A., 1996. Listeners do hear sounds, not tongues. J. Acoust. Soc. Am. 99, 1730–1741.

Fowler, C.A., 2006. Compensation for coarticulation reflects gesture perception, not spectral contrast. Percept. Psychophys. 68 (2), 161–177.

Fowler, C.A., Best, C.T., McRoberts, G.W., 1990. Young infants' perception of liquid co-articulatory influences on following stop consonants. Percept. Psychophys. 48, 559–570.

Fowler, C.A., Dekle, D., 1991. Listening with eye and hand: crossmodal contributions to speech perception. J. Exp. Psychol.: Hum. Percept. Perform. 17, 816–828.

Fowler, C.A., Housum, J., 1987. Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. J. Mem. Lang. 26, 489–504.

Fowler, C., Smith, M., 1986. Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In: Perkell, J., Klatt, D. (Eds.), Invariance and variability of speech processes. Erlbaum, Hillsdale, NJ, pp. 123–136.

Galantucci, B., Fowler, C., Turvey, M.T., 2006. The motor theory of speech perception reviewed. Pyschonomic Bull. Rev. 13, 361–377.

Ganong, W.F., 1980. Phonetic categorization in auditory word recognition. J. Exp. Psychol. Hum. Percept. Perform. 6, 110–125.

Gibson, E.J., Pick, A.D., 2000. An Ecological Approach to Perceptual Learning and Development. Oxford University Press, New York, pp. 1–238.

Gibson, J.J., 1966. The Senses Considered as Perceptual System. Houghton-Mifflin, Boston, MA, pp. 1–335.

Goldstein, L., Fowler, C.A., 2003. Articulatory phonology: a phonology for public language use. In: Schiller, N.O., Meyer, A.S. (Eds.), Phonetics and Phonology in Language Comprehension and Production. De Gruyter Mouton, Berlin, Boston, pp. 159–207.

Hagiwara, R., 1995. Acoustic Realizations of American /r/ as Produced by Women and Men. UCLA Working Papers in Phonetics, vol. 90, pp. 1–187.

Harrington, J., Kleber, F., Reubold, U., 2008. Compensation for coarticulation, /u/-fronting, and sound change in standard southern british: an acoustic and perceptual study. J. Acoust. Soc. Am. 123 (5), 2825–2835.

Harrington, J., Kleber, F., Reubold, U., 2011. The contributions of the lips and the tongue to the diachronic fronting of high back vowels in standard British English. J. Int. Phon. Assoc. 41 (2), 137–156.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92 (1-2), 67–99.

Holt, L.L., Lotto, A.J., Kluender, K.R., 2000. Neighboring spectral content influences vowel identification. J. Acoust. Soc. Am. 108, 710–722.

Johnson, K., 2000. Adaptive dispersion in vowel perception. Phonetica 57, 181–188.

Johnson, K., 2011. Retroflex Versus Bunched [r] in Compensation for Coarticulation. Annual Report. UC Berkeley Phonology Lab, pp. 114–127.

Johnson, K., 2012. Acoustic and Auditory Phonetics, third ed. Oxford: Wiley-Blackwell, pp. 1–232. (1st edition, 1997).

Johnson, K., Flemming, E., Wright, R., 1993. The hyperspace effect: Phonetic targets are hyperarticulated. Language 69, 505–528.

Johnson, K., Bakst, S., 2014. Audio/visual correlates of a vowel near-merger in the bay area. J. Acoust. Soc. Am. 135 (4), 2423.

Kataoka, R., 2011. Phonetic and Cognitive Bases of Sound Change. (Doctoral dissertation). University of California at Berkeley.

Klatt, D.H., 1980. Software for a cascade/parellel formant synthesizer. J. Acoust. Soc. Am. 67, 971–995.

Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., Lindblom, B., 1992. Linguistic experiences alter phonetic perception in infants by 6 months of age. Science 255, 606–608.

Lewkowicz, D.J., Minar, N.J., Tift, A.H., Brandon, M., 2015. Perception of the multisensory coherence of fluent audiovisual speech in infancy: its emergence and the role of experience. J. Exp. Child Psychol. 130, 147–162.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36.

Linker, W., 1982. Articulatory and Acoustic Correlates of Labial Activity in Vowels: A Cross-Linguistics Study. UCLA Working Papers in Phonetics 56, 1–134.

Lisker, L., Rossi, M., 1992. Auditory and visual cueing of the [+/- rounded] feature of vowels. Lang. Speech. 35 (4), 391–417.

Lotto, A.J., Kluender, K.R., 1998. General constrast effects in speech perception: effect of preceding liquid on stop consonant identification. Percept. Psychophys. 60, 602–619.

Mann, V.A., 1980. Influence of preceding liquid on stop-consonant perception. Percept. Psychophys. 28, 407–412.

Mann, V.A., 1986. Distinguishing universal and language-dependent levels of speech perception: evidence from japanese listeners perception of english "l" and "r". Cognition 24, 169–196.

Mann, V.A., Repp, B.H., 1980. Influence of vocalic ontext on perception of the [ʃ]-[s] distinction. Percept. Psychophys. 28 (3), 213–228.

Mann, V.A., Repp, B.H., 1981. Influence of preceding fricative on stop consonant perception. J. Acoust. Soc. Am. 69, 548–558.

McGurk, H., Macdonald, J., 1976. Hearing lips and seeing voices. Nature 264 (5588), 746–748.

McQueen, J.M., et al., 2006. Are there really interactive speech processes in speech perception? Trends Cogn. Sci. 10, 533.

McQueen, J.M., Jesse, A., Norris, D., Aubin, J., 2009. No lexical–prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? J. Memory and Language 61, 1–18. doi:10.1016/j.jml.2009.03.002.

Ménard, L., Schwartz, J.L., Boë, L.J., Aubin, J., 2007. Articulatory-acoustic relationships during vocal tract growth for french vowels: analysis of real data and simulations with an articulatory model. J. Phonetics 35, 1–19.

Mitterer, H., 2006. On the causes of compensation for coarticulation: evidence for phonological mediation. Percept. Psychophys. 68 (7), 1227–1240.

Noiray, A., Cathiard, M.A., Ménard, L., Abry, C., 2011. Test of the movement expansion model: anticipatory lip protrusion and constriction in french and english speakers. J. Acoust. Soc. Am. 129 (1), 340–349.

Nygaard, L.C., Eimas, P.D., 1990. A new version of duplex perception: evidence for phonetic and nonphonetic fusion. J. Acoust. Soc. Am. 88 (1), 75–86.

Pastore, R.E., Farrington, S.M., 1996. Measuring the difference limen for identification of order of onset for complex auditory stimuli. Percept. Psychophys. 58 (4), 510–526.

Pinheiro, J.C., Bates, D.M., 2004. Mixed-Effects Models in S and S-PLUS. Springer, New York, NY.

Pitt, M.A., McQueen, J.M., 1998. Is compensation for coarticulation mediated by the lexicon? J. Mem. Lang. 39, 347–370.

Pitt, M.A., Samuel, A.G., 1995. Lexical and sublexical feedback in auditory word recognition. Cogn. Psychol. 29 (2), 149–188.

Reed, E., 1996. Encountering the World: Toward an Ecological Psychology. Oxford University Press, Oxford.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. Science 212, 947–950.

Satish K.S., 2012. Wax [Computer Software]. Ver. 2.02. http://www.debugmode.com/wax/ (date last viewed 07/31/14).

Smits, R., 2001. Evidence for hierarchical categorization of coarticulated phonemes. J. Exp. Psychol.: Hum. Percept. Perform. 27, 1145–1162.

Strange, W., Levy, E.S., Law II, F.F., 2009. Cross-language categorization of french and german vowels by naïve American listeners. J. Acoust. Soc. Am. 126 (3), 1461–1476.

Strange, W., Weber, A., Levy, E.S., Shafiro, V., Hisagi, M., Nishi, K., 2007. Acoustic variability within and across german, french and american english vowels: phonetic context effects. J. Acoust. Soc. Am. 122 (2), 1111–1129.

Traunmüller, H., Öhrström, N., 2007. Audiovisual perception of openness and lip rounding in front vowels. J. Phon. 35, 244–258.

Viswanathan, N., Magnuson, J.S., Fowler, C.A., 2010. Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. J. Exp. Psychol. Hum. Percept. Perform. 36 (4), 1005–1015.

Viswanathan, N., Magnuson, J.S., Fowler, C.A., 2013. Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects in compensation for coarticulation. J. Exp. Psychol.: Hum. Percept. Perform. 39 (4), 1181–1192.

Whalen, D.H., 1981. Effects of vocalic formant transitions and vowel quality on the english [s]-[ʃ] boundary. J. Acoust. Soc. Am. 69 (1), 275–282.

Whalen, D.H., Liberman, A.M., 1987. Speech perception takes precedence over nonspeech perception. Science 237 (4811), 169–171.

Yang, B., 1986. A comparative study of American English and Korean vowels produced by male and female speakers. J. Phonetics 24, 245–261.