Effects of native-language on compensation for coarticulation

Shinae Kang, Keith Johnson, Greg Finley*

Abstract

This paper investigates whether compensation for coarticulation (CfC) in speech perception can be mediated by linguistic experience. CfC has been attributed to auditory or gestural recovery processes, but little is known about the role of linguistic experience. In Experiment 1, French and English native listeners identified an initial sound from a set of fricative-vowel syllables on a [s]-to- [ʃ] continuum 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 CfC effect with the vowel [u], but only the French listeners compensated for the vowel [y]. In Experiment 2, twenty-four American English listeners saw the video in which the audio stimuli of Experiment 1 were used as the soundtracks of a face saying [s]V, [ʃ]V, or a visual-blend of the two fricatives. The result shows that [ʃ]- and blend-videos induce significantly more [ʃ]-responses than [s]-videos. Again as in Experiment 1, native English listeners compensated for [u], but not for the unfamiliar vowel [y]. Interestingly, the compensation effect was not strengthened by seeing the lip rounding of [y]. The results indicate that CfC is a language specific effect tied to the listener's experience with the conditioning phonetic environment.

I INTRODUCTION

General properties of the auditory system determine what can and cannot be heard, what 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

*This manuscript has been submitted to JASA and recieved revise and resubmit. The authors are currently (as of Dec 20., 2014) revising the manuscript. We would like to thank all members of Phonology lab for their suggestions.

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 msec (Pastore & Farrington 1996). In addition to these general factors, speech perception may also be shaped by the two aspects of the listeners' knowledge.

In the first of these, experience with producing the phonetic segments of speech provides listeners with a base of knowledge that makes available a phonetic mode of listening (or speech mode; Liberman & Mattingly 1985) that elaborates and reinterprets the auditory image of speech. For instance, signal components that might not ordinarily be grouped with each other (Bregman 1990) do cohere in speech perception (”duplex perception”; Whalen & Liberman 1987). The phonetic mode underlies the strong tendency to hear foreign speech in terms of native segments (Best 1995) and the tendency for multimodal information to be combined into a phonetic percept (McGurk & MacDonald 1986). Perception of sine wave analogs of speech shifts suddenly from nonphonetic to phonetic (Remez et al. 1981, 1994).

The second set of linguistic factors is centered around the fact that the listener's ultimate aim in speech communication is to figure out what words the speaker is saying. At this level there are numerous lexical effects in speech perception. For example, perceptual errors (”slips of the ear”) overwhelmingly result in words (Bond 2008). Similarly, the Ganong effect (Ganong 1980, Fox 1984) shows 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 (Warren 1970, Pitt & Samuel 1995), and the restored phones interact with phonetic mode processes like compensation for coarticulation (Elman & McClelland 1988; Magnuson et al. 2003).

Proponents of each of these three aspects of speech perception have argued to preclude effects from other aspects (e.g. Fowler 2006 against auditory speech processing; McQueen et al. 2006 against lexical involvement in speech perception; and Lotto & Kluender 1998 , Diehl et al. 1989 against a specifically phonetic mode of processing). However, recent findings from neural imaging studies (cf. Liebenthal et al. 2003; Hickock & Poepple 2004) indicate that all three have something to do with the human ability to perceive speech. Ultimately, in addition to uncovering the neural organization of speech perception, we will want to know what listening circumstances recruit greater or lesser phonetic processing, or lexical processing, and what aspects of speech perception ultimately derive more from auditory processing than from specifically linguistic processing.

In this paper, we explore how the phonetic mode of listening may be shaped by linguistic experience in a compensation for coarticulation (CfC) task.[1] Compensation for coarticulation (Mann 1980, Mann & Repp 1981) is a listeners' perceptual demodulation of coarticulatory information during speech perception. For example, Mann & Repp (1980) found that the lower fricative pole induced by adjacent vowel 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 an inherent property of the fricative itself is a prototypical case of compensation for coarticulation. CfC has been investigated in many studies of consonant-vowel interactions in consonant place perception (Mann & Repp, 1980; Whalen 1981, 1984, 1989; Johnson 1991; Smits 2001; Mitterer 2006), vowel perception (Lindblom & Studdert-Kennedy 1967; Holt, Lotto & Kluender, 2000), and consonant voicing perception (Diehl & Walsh 1989), as well as in vowel-vowel interactions (Fowler, 1981; Fowler & Smith 1986; Beddor et al. 2002), and in consonant-consonant interactions (Mann 1980, Mann & Repp 1981, Pitt & McQueen 1998, Lotto & Kluender, 1998, Samuel & Pitt 2003, Fowler 2006).

Much of this literature is steeped in controversy regarding the basis of the compensation mechanism—whether it is auditory interaction between adjacent segments, or compensation is an indication of the undoing of the gestural interactions inherent in speaking. While auditory frequency contrast has been suggested to play a primary role in the phenomenon of CfC (Lotto& Kluender 1998, Sitek 2011, Johnson 2011), thus in support of an auditory account, several studies have provided evidence supporting a gestural account. Mitterer's (2006) study of perceptual compensation concludes that CfC has a phonological basis because when listeners see vowel rounding in audio-visual stimuli, they produce a compensation response with stimuli that didn't invoke such response in audio-only presentation. Viswanathan et al. (2010) also found evidence in favor of a gestural account. They tested CfC with Tamil retroflexs and liquids where the gestural compensation for coarticulation and spectral contrast predict opposite patterns. They found the pattern predicted by the gestural account and concluded that the auditory account alone is not sufficient in explaining CfC. It is clear that both Mitterer (2006) and Viswanathan et al. (2010) assume that knowing the articulation of a sound leads to or comprises the phonological knowledge of the sound. In this study, we aim to see if gestural knowledge extends to foreign sounds that listeners have little experience with. Also, 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 (Miyawaki et al. 1975, Werker & Tees 1984, Best et al. 1988), only a few studies of cross-liguistic CfC have been reported. Hence, the cross-linguistic study reported here sheds light on the issue by exploring the language specificity of compensation for coarticulation in audio-only and audio-visual presentation modalities.

II EXPERIMENT 1

Experiment 1 is a cross-linguistic study of CfC in which we measure the linguistic basis of the phenomenon by comparing compensation effects for native and nonnative sounds. Smits (2001) used the high front round vowel [y] with Dutch-speaking listeners, while Mann & Repp (1981) used the high back round vowel [u] with English-speaking listeners. The present experiment use both of these round vowels as potential triggers of CfC. One group of listeners (French) have native language familiarity with both [y] and [u], while the other group (English) is only familiar with [u]. It has been found that phonetically trained listeners are not very good at detecting vowel rounding for unfamiliar vowels from acoustic signals alone (Lisker & Rossi 1992, Traunmüller & Öhrström, 2007). Ettlinger & Johnson (2011) also found that experience with a feature such as rounding did not translate to skill in dealing with that feature on 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 [round] vowels used in the experiment. If linguistic experience guides CfC, then we would expect American English-speaking listeners to show less compensation in the [y] context than French-speaking listeners do. The cross-linguistic design lets us compare French speakers as a control group because the degree of rounding coarticulation caused by [y] may be less than that that caused by [u]. Finding a difference in C4C between the [u] and [y] context for English listeners is not as informative as seeing a difference between French and English listeners in the CfC effect triggered by [y].

A Methods

1 Stimuli

In making the auditory stimuli, six single CV syllables consisting of a fricative (/s/ or /ʃ/) and a vowel (/a/, /u/, or /y/) were first recorded by a native German speaker ([sa], [su], [sy]/ [ʃa], [ʃu], [ʃy]) with a Canon Model XF 100a camcorder with high definition audio and video at a audio sampling rate of 44100 hz and 30 fps for video. The vowels /a/, /u/, and /y/ were extracted from the audio track of the recording. The three vowels from [sa], [su] and [sy] were segmented from the recording and saved as separate .wav files. The onset of voicing was considered to be the beginning of the vowel for this purpose. To prevent clicks, the vowels were also given a brief (50 ms) linear fade-in.

Fricatives were synthesized using the Klatt terminal analog synthesizer (Klatt 1980) based on the fricatives [s] and [ʃ] preceding vowel [a] (where the fricatives are naturally maximally different from each other). The synthetic fricatives were 240 ms in length and were adjusted to have amplitude matching the natural fricatives (See Appendix A for the full set of synthesis parameters and acoustic vowel measurements). The synthesized fricatives and the extracted natural vowels were then concatenated to produce CV syllables. Figure 1 shows an example.

....
insert Fig 1 here

....

The synthetic [s] and [ʃ] were used as the endpoints of a 9-step fricative continuum in which the pole frequencies and amplitude stepped gradually from [s] to those of [ʃ]. The frequency steps of the continuum were equally spaced on the bark frequency scale. Concatenating the continuum 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. Figure 2 shows the spectral slices of the nine-step fricative continuum.

....
insert Fig 2 here

....

2 Subjects

Forty-two listeners 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 twenty-one French listeners were recruited at Université Pierre-Mendès-France, Grenoble, France. French listeners were recruited in particular, because 1) vowel [y] is phonemic in French and 2) yet, French listeners would still hear some foreign aspect from the stimuli produced by a German speaker as American English listeners would. None of the subjects reported any problem with hearing ability. Several participants from 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.

3 Procedure

The 27 CV tokens (9-step continuum x 3 vowels) were iterated seven times and presented to the participants at a random order. The participants heard one CV-stimuli at a time ‘s’ or ‘sh’ (9x7x3=218 trials). The inter-trial interval was 1 second.

4 Statistical Analysis

In order to test language-specificity in perceptual compensation for coarticulation, the responses were analyzed in a mixed effects logistic regression The model was built with several predictors. VOWEL (/a/, /u/, /y/) indexed what vowels followed the frication noise. TOKEN (range: 1∼9) was treated as a continuous variable indexing the fricative noise from the most [s]-like to the most [ʃ]-like fricative respectively. LANGUAGE (English vs. French) indexed the native language of each listener. Finally, LISTENERS (42 levels) indexed each listener. The dependent variable was listeners’ RESPONSE (‘s’ vs. ‘sh’). “s”-responses were coded as 0 and “ʃ”-responses as 1, thus positive model coefficients indicate greater probability of “ʃ”-responses and negative coefficients indicate greater likelihood of “s”-responses.

Listener-specific variation in the probability of “s”-response was accounted for by including LISTENERS random intercepts. TOKEN was first included in the model to control for the effect of the frication noise itself on listeners response. In order to test the main interest of the Experiment, whether there is language-specific compensation to different vowels, LANGUAGE and its interactions with VOWEL were included in the model as fixed effects. Goodness of fit of the regression models was calculated by both marginal R^2 (R^2_m) and conditional R^2 (R^2_c), as a point estimate of the variance explained by a model, using the maximum likelihood estimates of the model parameters and quantifying the uncertainty around these estimates using Monte-Carlo Markov Chain (MCMC) sampling (Nakagawa and Schielzeth 2013). The significance of model improvement was compared using a likelihood-ratio test via Analysis of Deviance (cf. Baayen 2011). See also Appendix B for the results of an additional repeated measures ANOVA.

B Result

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 Figure 3.

....

insert Fig 3 here

....

The pattern in Figure 3 reflects how auditory continua were made. Since nine tokens were created by interpolating the synthesis parameters of the two endpoints, which represented the most [s]-like sound and the most [ʃ]-like sound respectively, the proportion of "s"-responses naturally decreases along the continua from Token 1 to Token 9. The response pattern differs by vowel environments: Round vowels elicit more "s"-response for both English and French listeners. More "s"-responses for the tokens following round vowels indicate that listeners compensate for the effects of rounding, the pattern is in line with previous findings (Mann &Repp 1980, Smits 2001, Mitterer 2006).

The regression model fit to the probability of responses indicates that Token itself accounts for over 70 percent of the variance in response (see table 1). Taking the model having only Token as a fixed effect as the basline, adding Vowel to the model increases the fit significantly [χ^2= 108.8, df=2, p<0.01]. Including Language as a main effect does not itself improve the model fit [χ^2=2.699, df=1, n.s.]. However, a model with the Language:Vowel interaction significantly improves the fit [χ^2= 18.239, df=2, p<0.001]. No other interaction term (Token*Vowel, Language*Token, Token*Vowel*Lg) was found to significantly improve the model fit. In Table 1, we report only a marginal R^2 (R^2_m), since the random intercept is constant across all models.

Table 2 shows the fixed-effect coefficients in a relatively full model (token +vowel +language + vowel:language) of the experiment 1 results. Negative estimates are associated with greater likelihood of "s"-responses and the positive estimates as greater likelihood of "ʃ"-response. As shown, there is a significant positive effect of token, suggesting thatResponse is more likely to be "sh" as Token number increases. Both round vowels [u] and [y] have significant negative coefficients: Responses are more likely to be "s" before the two round vowels, but with greater effect for [u] than for [y].

Although we did not find a significant effect of Language itself, there was a significant negative effect of Vowel and Language interaction. Figure 4 illustrates predicted proportions of "s"-responses for vowel [u] and [y] by American English and French listeners. As shown, French listeners are more likely to respond [s] before round vowels than English listeners are and the pattern holds valid for both [u] and [y]. The magnitude of the effect is again greater for vowel [u] than [y].

....
insert Fig 4 here

....

C Interim summary

The results of experiment 1 suggest that both English listeners and French listeners compensate for rounding. However the magnitude of the effect is larger for French listeners than it is for English listeners. There is no significant main effect of Language, which is partly attributable to the similarities in the pattern where [u] induced greater compensation effect than [y] did. However, the significant interaction of Language and Vowel suggests that the effect of rounding was different for French and American English listeners. The finding is that French listeners showed a stronger compensation effect for [u] context, and where English listeners showed no effect for [y] context, French listeners did. The fact that French has both [u] and [y] as native phonemes could have made it possible for French listeners to be more sensitive to rounding. On the other hand, only [u] holds a status of native phoneme in English. It is plausible therefore that for English speakers, [y] is not recognized as round, while for French speakers, [y] like [u] both produced a substantial compensation effect.

III EXPERIMENT 2

Experiment 2 is an audiovisual extension of Experiment 1. Traunmüller and Öhrström (2007) found that rounding is strongly signaled in visual displays. Also, Mitterer (2006) used visual lip rounding to induce a vowel rounding percept which produced the CfC reaction in fricative identification. If, as preliminary data suggest, the compensation for coarticulation effect is dependent upon linguistic experience, then the strongest test of this basis of CfC is to present both audio and visual lip rounding. If English-speaking listeners continue to be less sensitive to vowel rounding, and show no CfC effect in the [y] environment, then we would have to conclude that CfC is strongly mediated by linguistic experience. We may also find, though, that the compensation response is much stronger when the vowel rounding information in the audio/visual speech signal is stronger. This would indicate that there are both gestural and linguistic components in the CfC response.

A Methods

1 Stimuli

In order to see the effect of visual information (rounding), the same auditory stimuli as those used in Experiment 1 were aligned with the original videos of the face of the model speaker articulating CV-syllables. 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 addition, in order to test effects of visual fricative portions, 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] continua 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 (Sampath 2012). Dynamic morphing is analogous to making auditory continua by interpolating the acoustic measurements of [s] and [ʃ]. The visual pair that served as the reference points in dynamic morphing were always selected to match the vowels of the auditory stimuli such that [sa] and [ʃa] produced a blended [sa]/[ʃa] visual stimulus, and blended [su]/[ʃu] and [sy]/[ʃy] were also created. The difference in the lip gestures during [s] and [ʃ] were not as conspicuous in fricatives before [u] and [y] as before the unround vowel [a]. So, all three visual fricative stimuli before round vowels were similar to one another as compared to those before non-round [a] vowel.

2 Subjects

Twenty-four listeners without any 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 Procedure

The participants saw the stimulus movies on a computer monitor at a distance of about 20 inches and heard the audio portion over headphones at a comfortable listening level. The subjects werew asked to identify the initial consonant as either "s" or "ʃ". 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 tokens x 3 vowels x 3 visual fricatives x 7 repetitions). The list of tokens was randomized and presented to the listener one at a time. The stimuli were divided into two blocks and the participants could take a short break between the two blocks.

4 Statistical Analysis

The analysis method used for Experiment 1 was also adopted for this experiment. In order to see the effect of visual fricative, the responses were analyzed with mixed effects logistic regression fitted in in R with a predictor variable, Visual Fricative (VF) (s, sh, s_sh), in addition to the predictors included in the model in Experiment 1: VOWEL ([a], [u], [y]), TOKEN (range: 2∼8) , and the random effect LISTENER (24 levels) . The dependent variable was listeners' RESPONSE ("s" vs. "ʃ"). "s"-Responses were coded as 0 and "ʃ"-responses as 1, thus positive coefficient of model slope indicate greater probability of "ʃ"-responses and negative weight greater chances of "s"-responses.

The goodness of fit of the regression models was calculated in the same way as Experiment 1. Listener-specific variation in the probability of "s" response was accounted for by including by-LISTENER random intercepts. Main effect terms for VOWEL, TOKEN, VISUAL FRICATIVE, as well as their interactions, were included in the models as fixed effects to test their effects on RESPONSE. Goodness of fit of the regression models was calculated by both marginal R^2 (R^2_m) and conditional R^2 (R^2_c), as a point estimate of the variance explained by a model, using the maximum likelihood estimates of the model parameters and quantifying the uncertainty around these estimates using Monte-Carlo Markov Chain (MCMC) sampling (Nakagawa and Schielzeth 2013). The significance of model improvement between two models was compared using likelihood-ratio test via Analysis of Deviance (cf. Baayen 2011).

B Results

1 Overall results

The mean proportion of "s" responses across all subjects for each fricative token in the three different vowel environments is plotted in Figure 5. The three different visual fricative conditions are collapsed in Figure 5 and vowel context refers to both the audio vowel and the matching visual vowel - in this experiment audio and visual vowels always matched.

....
insert Fig 5 here

....

As in Experiment 1, the unround vowel /a/ yields 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 the results seem to show a compensation effect at least with the vowel /u/ with a weak or non-existent compensation effect for the vowel /y/.

Regression models indicate that TOKEN can account for approximately 70 percent of the variance of response (R^2_GLMM$(m) = 0.716$,R^2_GLMM$(c) = 0.802$m see Table 3). Adding the fixed terms significantly improves the mean estimates of the model: Inclusion of Vowel improves R^2_GLMM(m) to 0.726 and conditional R^2_GLMM(c) to 0.813 (Analysis of Deviance: χ^2= 76.227, df=2, p<0.01 **). Although Visual Fricative alone does not improve the model fit, to allow the possibility that visual fricative has differential effect on Response before round vowel [u] /[y] and before unround vowel [a], VF, Vowel, and their interactions were included as the fixed terms. As a result, adding Visual Fricative and VF and Vowel interaction improves the model even further to 0.728 for R^2_GLMM(m) and to .0813 forR^2_GLMM(c) (Analysis of Deviance: χ^2= 14.268, df=6, p<0.05 *). No other interaction terms (Token * Vowel, VF*Token, Token*Vowel*VF) was found to significantly improve the data likelihood (p>0.1) (summarized in Table3).

The estimated values for all fixed-effect predictors in the full model are listed in Table 4.

The way we coded the response allows the negative estimates to be interpreted as more likelihood of "s"-responses and the positive estimates as greater likelihood of "ʃ"-response. Like Experiment 1, mean estimates in Table 4 indicate a significant positive effect of Token: Responses are more likely to be "ʃ" as Token number increases. Both round vowels [u] and [y] have significant negative effects on Responses: Responses are more likely to be "s" before the two round vowels, but with greater magnitude for [u].

Of particular interest in this experiment condition is the effect of Visual fricative on response, which serves as a way to see if the listener is affected by visual cues at all. There is a significant positive effect of Visual Fricative. Visual [ʃ] led to a greater number of "sh" responses than did visual [s], and the ambiguous visual [s_sh] led to a marginally (p=0.054) greater number of "ʃ"-responses. The result is plausible since the three visual fricatives constitute a visual continuum where visual "s" gradually shifts to visual "sh". Figure 6 shows the predicted proportions of "s" responses in the three visual fricative conditions.

....
insert Fig 6 here

....

The only coefficient in the Token:Visual Fricative interaction to reach significance was for the combination of [u] with visual fricative [ʃ]. Lack of the similar effect in vowel [y] might be attributable to the possibility that [y] is an unfamiliar vowel so that even when the gesture is explicitly presented to the listener, the listeners do not use the visual evidence of lip rounding in a compensation effect.

Post-hoc pairwise comparisons between every vowel pair indicate a significant difference in the response between vowel /a/ and /u/ and the response between vowel /a/ and /y/, but not with /u/ and /y/. The result is slightly different from Experiment 1, in that unlike Experiment 1 the compensation effect is found in both round vowels. The result in Experiment 2 suggests that American English listeners compensated for round vowel with visual presenation. However, it is yet unclear whether the compensation indeed improved as a result of visual vowel, or the listeners in Experiment 2 were overall more sensitive than the listeners in Expeirment 1. After all, the compensation for vowel /y/ was much smaller even for French listener, which indicates that the degree to which listeners are sensitive to rounding is more important than the mere presence or absence of the compensation effect. Due to the lack of the French listener group as a control, we cannot conclude whether or not the English listeners in Experiment 2 became as sensitive as the French listenrers group.

2 Effects of Visual vowel

Instead, to further test whether seeing the lip rounding during articulation indeed affects American English listeners' compensation to round vowels overall, responses from Experiment 1 (audio-only modality) and Experiment 2 (audiovisual modality) were combined and analyzed together in a mixed-effects logistic regression. The model here contained Modality(Audio vs. Audiovisual) as an additional predictor along with other predictors mentioned previously (Token, Vowel). For this analysis we used only responses to the blend fricative ('s_sh') in order to minimize any effect of visual fricative. We expected that listeners would rely more on the visual vowel when the visual fricative is ambiguous. Listener-specific variation in the probability of "s"-response was accounted for by including by-Listener random intercepts. Main effect terms for Vowel, Token, Modality as well as their interactions were included in the model as fixed effects to test their effects on Response.

Regression models fit to the probability of "s"-responses indicate that Vowel and Token together account for nearly 80 percent of the variance in response (R^2_GLMM(m) = 0.739, R^2_GLMM(c) = 0.812). Modality and the Modality:Vowel interaction were added to the model as fixed-effects terms, but failed to make any significant improvement to the model fit (χ^2=2.37, n.s.). All other interactions (Token*Modality, Modality*Vowel, Token*Modality*Vowel) did not improve the model fit. See Table 5 for full estimates for the model. Thus, the finding is that visual fricative information did influence listeners' identification of fricatives, while visual vowel rounding information did not add to the compensation effect (for [u] and not for [y]) that was otained with audio-only sitmuli.

IV General discussion

A Language-specific compensation for coarticulation

We found that participants heard more "s" in front of rounded vowels and more "sh" in front of 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. Vowel [u] consistently yielded more "s"-responses across all listeners regardless of their native language background. It is partly attributable to the inherent acoustics of [u] in which the effect of rounding (lowering of formants etc.) is more salient than [y]. According to the spectral contrast view of CfC, the contrast in the formants between fricative and vowel is what triggers compensation where listeners register a different auditory fricative depnding on the neighboring vowel. Hence, the more noticeable the contrast is, the more likely listeners are to compensate. In this sense, since [u] has lower F2 than [y], the contrast between vowel and fricative is greater for [u], which could have triggered CfC to the greater degree.

Although [u] is universally more susceptible to CfC due to its physical nature, we found an effect of the listener's native language on CfC. In Experiment 1, we found that the French listeners had a stronger compensation responses for [u] and [y] than did American English listeners. Since [y] is not a native phoneme of English, American English listeners are familiar only with [u], whereas French listeners whose native language contains /y/ in its phoneme inventory are familiar to both [u] and [y]. The French listeners' greater compensation effect might be attributable to the role of rounding in the French vocalic inventory. Having both [u] and [y] contrastive, French listeners must rely on rounding to distinguish [y] from [i], which shares other features with [y] such as place and height. On the other hand, [round] is redundant with [back] in English. The round vowels in English instead can be descibed 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. The result is taken as evidence that familiarity to phoneme can affect listeners' ability to compensate.

The result in Experiment 2 further supports the possibility that compensation is language-specific. Although the participants saw the explicit visual cue in which they could see the speaker rounding her lips as she produced [y], the number of "s"-responses by American English listeners did not increase significantly. If the source of the compensation effect had been the visual information as might be suggested by a gesture-perception account of CfC, compensation in the [y] context should have increased with addition of videos. Therefore, the result in this study supports the hypothesis that the compensation effect is strongly influenced by listeners' experience and that this cannot be overcome easily by seeing how the unfamiliar sound is articulated.

B Role of visual information in compensation for coarticulation

We tested whether CfC extends to unfamiliar sounds when visual information is presented. We did not find any increase of compensation with videos of either [u] or [y] in American listeners CfC, contra Mitterer (2006) where Dutch listeners were found to compensate more in audiovisual presentation. One possible explanation for the lack of a visual effect even on the native [u] phoneme can be found from the nature of the stimuli. While we concatenated synthesized fricative and natural vowels, the stimuli used in Mitterer (2006) contained three synthesized vowels that were selected from vowel continua (analogues to the fricative continua we used here). The synthetic vowels were quite ambiguous, which may have caused the size of the audio-only effect to be small. Hence it was possible in Mitterer's study to increase magnitude of CfC with the presence of another source of information, visual rounding. The natural vowels used in our study, on the other hand, had little room for such increase.

The effect of visual fricative in our data shows that the listener did pay sufficient attention to the task. If the listeners did not pay attention to the visual information, the effect of visual information should have been minimal. However, there is 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 hear "s" when they saw the /s/- visual fricative and more "sh"s when they saw /s_ʃ/- visual fricative, and even more so for /ʃ/- visual fricative. This pattern showing the effect of the visual fricative was not present in other vowel conditions where anticipatory vowel lip rounding reduced the visual differnce between /s/ and /ʃ/. The result that the participants response pattern matches the visual fricative input suggests that the participants were attending to, and using the visual input in this experiment. The result also may imply that processing of visual and audio information is automatic and unconscious, even when there was no certain adverse condition to trigger the process as in Mitterer (2006). Therefore, lack of visual advantage on CfC for [y] supports our hypothesis that CfC is also driven by more global knowledge such as native language and cannot always be modulated by mere gestural perception.

V CONCLUSION

We have shown that compensation for coarticulation is language-specific and visual perception of speech cannot itself change listeners' pattern of compensation. Also, we have found that language-specificity in CfC may be manifested by differential sensitivity to a feature depending on its phonological status in the given language. The findings suggest that CfC is a phenomenon that is modulated not only by sensory factors like spectral contrasts between segments or the specific gestures associated with segments, but also by phonetic knowledge of one's native language. Our result on audiovisual modality raises an important question about the assumption that knowing articulatory gestures is directly linked to the knowledge of sounds. It is possible that this assumption is only applicable to the sounds of one's native language. Since our result does not offer a conclusive answer to this question, further research with careful modulation with the stimuli may be necessary to reveal how and which visual properties American English listeners capture during CfC. Finally, our result with cross-linguistic comparisons sheds light on the role of linguistic experience in shaping linguistic/phonetic categories in spoken language processing.

Acknowledgements

We are grateful to the members of Phonology Lab at UC Berkeley Linguistics Department for their valuable suggestions. We also presented an earlier versions of this project at the 164th ASA meetings at Kansas City; we thank participants there for their feedbank. We woud like to give special thanks to Elsa Spinelli for help in data collection in France, Nicholas Strzelczyk and Carson Miller Rigoli for their work in helping us create the stimuli.

Appendix A

This appendix lists synthesis control parameters for the synthetic fricative continuum and reports acoustic phonetic measurements taken from the stimuli.

		[s]								[ʃ]
	Step	1	2	3	4	5	6	7	8	9
Overall Gain	g0	66	64	62	61	59	57	56	54	53
Amplitudes of Formants	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	5876	5775	5677	5581	5487	5394	5303	5213	5125

Table A.1: Synthesis parameters

synthesized fricative		center of gravity	std	skewness	kurtosis
s_sh_1	[s]	9271.6	1088.84	-0.74	6.96
s_sh_2		8775.94	1176.67	-0.82	3.25
s_sh_3		8294.42	1264.85	-0.73	2.49
s_sh_4		7946.06	1343.55	-0.89	3.23
s_sh_5		7575.3	1524.27	-0.61	4.04
s_sh_6		6967.15	1754.36	-0.61	1.29
s_sh_7		6086.75	1993.21	-0.17	-0.43
s_sh_8		5467.93	1972.98	0.14	-0.87
s_sh_9	[ʃ]	4693.03	1984.78	0.81	1.16

Table A.2: Spectral moments analysis of the fricative continuum

Vowel	**a**	**u**	**y**
F1	801	432	427
F2	1572	1217	2331
F3	3087	3208	3046

Table A.3: Vowel formant analysis of the stimulus vowels.

Appendix B

In addition to mixed-effects modeling, we also analyzed the data using repeated measures ANOVA. The dependent measure is proportion of "s" responses for each listener in each condition. Post-hoc pairwise comparisons were done with paired t-test with adjusted p-value.

Experiment 1. For experiment 1 (table 9), the two-way repeated measures ANOVA is consistent with the mixed effects model reported in the text. The ANOVA found a main effect of Vowel and a Language*Vowel interaction. Post-hoc pairwise comparisons with adjusted level of significance reveal that English listeners showed compensation only for [u] while French listeners compensated for both [u] and [y].

	Df	**Sum Sq.**	**Mean Sq.**	***F***	**p**
Language	1	0.076	0.0756	3.45	0.07 *
Vowel	2	0.182	0.091	30.804	<0.01***
Lang*Vowel	2	0.249	0.0125	4.222	0.01 **

Table A.4: Two-way Repeated measure ANOVA for Experiment 1.

Experiment 2. Analysis using repeated measures ANOVA is consistent with the mixed effects model reported in the text. There were significant main effects of vowel environment, and visual fricative, and the interaction of the two was also reliable (Table 10). Pairwise comparisons of the three visual fricative conditions by vowel indicates that the effect was most reliable with the unround vowel /a/. The round vowels mostly yielded similar response patterns across all visual fricative conditions. Descriptive statistics are presented in table 11. In short, the participants were not greatly affected by which fricative they saw, except in the context of /a/. This makes sense because with coarticulatory rounding both /s/ and /ʃ/ have rounded lips when they precede /u/ and /y/.

	Df	Sum Sq.	Mean Sq.	F	p
Visual Fricative	2	0.079	0.03948	3.839	0.0287 *
Vowel	2	1.19	0.5948	6.446	0.00341 **
VF*Vowel	4	0.1124	0.028102	3.029	0.0215 *

Table A.5: Two-way repeated measures ANOVA results for Experiment 2.

		N	Mean	SD	SE mean
s	a	24	0.47	0.13	0.03
	u	24	0.52	0.11	0.02
	y	24	0.49	0.11	0.02
sh	a	24	0.43	0.12	0.02
	u	24	0.53	0.12	0.02
	y	24	0.48	0.12	0.02
s_sh	a	24	0.45	0.11	0.02
	u	24	0.52	0.1	0.02
	y	24	0.47	0.09	0.02

Table A.6: Descriptive statistics for experiment 2. Mean proportion of s responses and the Standard Deviation and Standard Error of the mean as a function of visual fricative, and vocalic context.

Notes

[1] It should be noted that our focus on the phonetic mode does not imply that we discount auditory and lexical factors in speech perception. Our experiments on CfC 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 debating whether a lexically biased percept can induce compensation for coarticulation (Ellman & McClelland 1988; Magnuson, et al. 2003, Pitt & McQueen 1998, Samuel & Pitt 2003), 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.

References

Baayen, R.H. (2011) languageR. R package .Available: http://cran.r-project.org/web/packages/languageR/index.html (date last viewed 06/22/14).

Bates, D., Maechler, M., Bolker, B. (2011) lme4. R package version 0.999375-38 (date last viewed 06/22/14).

Beddor, P.S., Harnsberger, J.D. & Lindemann, S. (2002) Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics. 30*, pp. 591–627.

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 perspective on cross-language speech perception. In *Speech perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research*, edited by W. Strange (York: Timonium, MD), 167-200.

Boersma, Paul & Weenink, David (2014) Praat: doing phonetics by computer [Computer program]. Version 5.3.65, retrieved 27 February 2014 from http://www.praat.org/ (date last viewed 06/22/14).

Bond, Z. S. (2005) "Slips of the ear" in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford: Blackwell, pp. 290–310.

Bregman, A.S. (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, Massachusetts: The MIT Press. 1–792.

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. *Journal of Memory and Language 27*(2), 143–165.

Fowler, C.A. (1981) Production and perception of coarticulation amon stressed and unstressed vowels. *J. Sp. & Hear. Res. 46*, 127–139.

Fowler, C.A. & Smith, M. (1986) Speech perception as vector analysis: an approach to the problems of segmentation and invariance. In J.S. Perkell & D.H. Klatt (Eds.), *Invariance and variability of speech processes*. Hillsdale, NJ: Erlbaum. 123–139.

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

Fox, R.A. (1984) Effect of Lexical Status on Phonetic Categorization. *J. Exp. Psychol. Hum. Percept. Perform. 10*, 526–540.

Ganong, W.F. (1980) Phonetic Categorization in Auditory Word Recognition. *J. of Exp. Psychol. Hum. Percept.Perform. 6*, 110–125.

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. (1991) Differential effects of speaker and vowel variability on fricative perception. *Language and Speech 34*, 265–279.

Johnson, K. (2011) *Acoustic and Auditory Phonetics.* 3rd Edition (1st edition, 1997). Oxford: Wiley-Blackwell. 1–232.

Johnson, K. (2011) Retroflex versus bunched [r] in compensation for coarticulation *UC Berkeley Phonology Lab Annual Report*, 114–127.

Klatt, D. H. (1980) Software for a cascade/parellel formant synthesizer. *Journal of the Acoustical Society of America. 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.

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

Liebenthal, E., Binder, J. R., Piorkowski, R. L., &, R. E. (2003). Short-term reorganization of auditory analysis induced by phonetic experience. *Journal of cognitive neuroscience 15*(4), 549–558.

Lindblom, B. E. & Studdert-Kennedy, M. (1967) On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am. 42*, 830–843.

Lisker, L. & Rossi, M. (1992) Auditory and visual cueing of the [+/- rounded] feature of vowels. *Language & 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.

Magnuson, J.S., McMuray, B., Tanenhaus, M.K. & Aslin, R.N. (2003) Lexical effects on compensation for coarticulation: the ghost of Christmash past. *Cognitive Science 27*, 285–298.

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

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. doi:10.1038/264746a0

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

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics 68* (7), 1227–1240.

Miyawaki, K., W. Strange, R. Verbrugge, A. M. Lieberman, J. J. Jenkens & O. Fujimura (1975) An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics 18*, 331–340.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. (R. B. OHara, Ed.) *Methods in Ecology and Evolution 4* (2), 133–142.

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

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. *Cognitive psychology 29* (2), 149–188.

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

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological review, 101* (1), 129–56.

Samuel, A.G. & Pitt, M.A. (2003) Lexical activation (and other factors) can mediate compensation for coarticulation. *J. Mem. & Lang. 48*, 416–434.

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

Sitek, K. (2011) Ipsilateral and contralateral phonetic context effects. *UC Berkeley Phonology Lab Annual Report*, 77–97.

Smits, R. (2001) Evidence for hierarchical categorization of coarticulated phonemes. *J. Exp. Psych.: H. Perc. Perf. 27*, 1145–1162.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of experimental psychology. Human perception and performance, 36*(4), 1005–1015.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science, 167*(3917), 392–393. doi:10.1126/science.167.3917.392.

Werker, J.F. & Tees, R.C. (1984) Phonemic and phonetic factors in adult cross-language speech perception. *J. Acoust. Soc. Am. 75*, 1866–1878.

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. (1984) Subcategorical mismatches slow phonetic judgements. *Percept. & Psychophys. 35*(1), 49–64.

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

Whalen, D.H. (1989) Vowel and consonant judgements are not independent when cued by the same information. *Percept. & Psychophys. 46*(3), 284–292.

Table 1: Experiment 1 models and fit statistics. R^2_m is the marginalR^2, the variance explained by the fixed factors in the model. χ^2 values are direct comparions of the model with the immediately preceding nested one.

		R^2_m	χ^2	p-value
baseline model	Token	0.723	—	—
+ Vowel	Token + Vowel	0.733	108.8	<**0.01**
+ L + Vowel:Language	Token + Vowel + L + V:L	0.740	18.239	<**0.01**

Table 2: Experiment 1: Mean estimates of mixed effect logistic model. Token, Language, Vowel, and Vowel:Language interaction are included as fixed-effects terms.

	β Estimate	Std.Error	z value	P-value
(intercept)	-6.612	.0284	-23.228	<**0.001**
Token	1.345	0.030	44.512	<**0.001**
French	-0.128	0.343	-0.373	0.709
[u]	-0.600	0.136	-4.396	<**0.001**
[y]	-0.305	0.136	-2.239	**0.025**
French:[u]	-0.822	0.193	-4.257	<**0.001**
French: [y]	-0.423	0.192	-2.203	**0.028**

Table 3: Experiment 2 models and fit statistics. R^2_m is the marginalR^2, the variance explained by the fixed factors in the model. χ^2 values are direct comparions of the model with the immediately preceding nested one.

		R^2_m	χ^2	p-value
baseline model	Token	0.716	—	—
+ Vowel	Token + Vowel	0.726	76.227	<**0.01***
+ Vowel + VF + Vowel:VF	T + V+VF + V:VF	0.728	14.268	<**0.05***

Table 4: 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.

	β Estimate	Std.Error	z value	P-value
(intercept)	-8.558	0.362	-23.661	<**0.001***
Token	1.775	0.146	38.469	<**0.001***
[u]	-0.633	0.194	-3.270	<**0.01***
[y]	-0.447	0.193	-2.312	**0.021***
vf(s_sh)	0.373	0.193	1.929	**0.054.**
vf(sh)	0.466	0.193	2.410	**0.016***
[u]: vf(s_sh)	-0.354	0.273	-1.297	0.195
[y]: vf(s_sh)	0.093	0.273	0.340	0.734
[u]:vf(sh)	-.0690	0.274	-2.521	**0.012***
[y]:vf(sh)	-0.224	0.273	-0.821	0.412

Table 5: Mean estimates of Mixed effects logistic model. Token, Vowel, Modality (Mode), and Vowel:Mode interaction are included as fixed-effects terms. Refer to the body for the full descriptions of the model.

	β Estimate	Std.Error	z value	P-value
(intercept)	-7.342	0.326	-22.543	<**0.001***
TOKEN	1.494	0.039	38.509	<**0.001***
[u]	-0.664	0.144	-4.613	<**0.001***
[y]	-0.337	0.143	-2.354	**0.019***
AUDIOVISUAL(AV)	0.459	0.371	1.236	0.216
[u]:AUDIOVISUAL(AV)	-0.174	0.228	-0.763	0.446
[y]:AUDIOVISUAL(AV)	0.036	0.228	0.129	0.874

A List of Figures