Journal of Phonetics 41 (2013) 145–155

Contents lists available at SciVerse ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Compensation for vocal tract characteristics across native and non-native languages

Matthias J. Sjerps [a,*], Rajka Smiljanić [b]

[a] Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[b] The University of Texas at Austin, Department of Linguistics, USA

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form
15 January 2013
Accepted 26 January 2013

ABSTRACT

Perceptual compensation for speaker vocal tract properties was investigated in four groups of listeners: native speakers of English and native speakers of Dutch, native speakers of Spanish with low proficiency in English, and Spanish–English bilinguals. Listeners categorized targets on a [sofo] to [sufu] continuum. Targets were preceded by sentences that were manipulated to have either a high or a low F_1 contour. All listeners performed the categorization task for targets that were preceded by Spanish, English and Dutch precursors. Results show that listeners from each of the four language backgrounds compensate for speaker vocal tract properties regardless of language-specific vowel inventory properties. Listeners also compensate when they listen to stimuli in another language. The results suggest that patterns of compensation are mainly determined by auditory properties of precursor sentences.

1. Introduction

Across-speaker variation in speech production due to vocal tract differences has been well documented (Peterson & Barney, 1952). These differences contribute to variation in the exact acoustic–phonetic properties of speech sounds. For instance, a speaker with a long vocal tract will have a lower first formant (F_1) when uttering the vowel /o/ relative to a speaker with a short vocal tract. Variation due to differences in the size and shape of speakers' vocal tracts can be problematic for listeners, because formant frequencies are an important determinant of vowel identity. In principle, this would suggest that a listener would be inclined to misperceive an intended /o/ for /u/ (which has a lower F_1 in general) when listening to a speaker with a longer vocal tract.

Listeners, however, can use additional information in the speech signal to overcome potential misperceptions. Such information consists of, for example, both the higher formants and the speaker's pitch (Johnson, 2005; Nearey, 1989). This type of compensation has been termed intrinsic normalization because the necessary information is available within the target vowel itself (Nearey, 1989). In addition, listeners can utilize spectral information available in the context, that is, outside the target vowels, to guide their perception of vowels. Ladefoged and Broadbent (1957) showed that when listeners were categorizing sounds halfway between [ε] and [ɪ] (which have a high and a low F_1 respectively), they heard more sounds as /ɪ/ when the target word was preceded by a sentence with a high F_1 than with a low F_1. This showed that listeners compensated for the vocal tract properties of a speaker as revealed in a preceding sentence. It is unclear, however, whether such effects are largely caused by general auditory processes (and as such are mainly dependent on signal properties) or by a listener's compensation for phonetic properties of a specific speaker's vowel space. The latter suggests that these compensation effects are for an important part dependent on a listener's experience with auditory properties of phonetic categories in his or her native language. To address this possibility, the current paper investigated whether the perceptual compensation of vowels is influenced by a listener's native language.

Previous research has demonstrated that the direction and amount of perceptual compensation is in part determined by the relation of the Long Term Average Spectrum (LTAS) between a precursor sentence and a target sound (Kiefte & Kluender, 2008; Kluender, Coady, & Kiefte, 2003; Laing, Liu, Lotto, & Holt, 2012; Watkins, 1991; Watkins & Makin, 1994, 1996). These authors have reasoned that a precursor sentence with high amplitudes in high spectral regions will suppress the perceptual impact of those frequencies in a following target (and, similarly, for precursor sentences with high amplitudes in low spectral regions), causing contrastive influences of contexts on target perception. Such a contrastive mechanism, in fact, is known to play an important role when context and target stimuli are presented with very short or no

[*] Corresponding author. Tel.: +31 641 857774.
E-mail addresses: m.j.sjerps@gmail.com, matthias.sjerps@mpi.nl (M.J. Sjerps).

0095-4470/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.01.005

interstimulus intervals. In such cases it has been argued that these effects partly arise in the peripheral system (Summerfield, Haggard, Foster, & Gray, 1984; Watkins, 1991; Wilson, 1970). However, similar compensation effects have been found over longer precursor–target intervals and with contralateral stimulation (Holt & Lotto, 2002; Lotto, Sullivan, & Holt, 2003; Sjerps, Mitterer, & McQueen, 2012; Watkins, 1991) even though these manipulations reduce peripheral effects to a minimum or even abolish them (Summerfield et al., 1984). The latter findings therefore suggest that an important part of compensation effects have a central origin (Holt & Lotto, 2002; Laing et al., 2012; Lotto et al., 2003; Watkins, 1991).

In spite of their central locus these compensation processes may still operate in a rather general auditory way (Holt & Lotto, 2002). In speech perception, when listeners categorize targets on, for example, an [ɪ] to [ɛ] continuum, a precursor with a relatively high F_1 might decrease the perceptual impact of higher frequencies in the subsequent targets. An ambiguous target sound that is preceded by a precursor with a high F_1 will then be perceived as one with a relatively low F_1 (and thus more /ɪ/-like) compared to the context in which the precursor sentence has a relatively low F_1. Perceptual compensation, then, acts in a fashion similar to applying a filter whose frequency response is the inverse of the LTAS of the precursor, to a target signal (Watkins & Makin, 1994, 1996). Mitterer (2006) has shown that when targets are preceded by a precursor sentence with a manipulated F_2 contour (either a high F_2 or a low F_2), effects on categorization of the targets are found only for vowels that have their formants in the same region as the manipulated formant in the precursor (in this case in the mid-to-high front vowel region). This suggests that central compensation effects are indeed largely determined by the spectral relation between a precursor and a target signal (see Laing et al., 2012, for a similar conclusion with consonant targets). This auditory approach therefore focuses on general signal properties that are independent of listener's background language and even the speech status of stimuli (Holt, 2005, 2006; Stilp, Alexander, Kiefte, & Kluender, 2010).

A general auditory approach contrasts with the framework that has been termed ''extrinsic vowel normalization'' (Nearey, 1989). According to this approach, listeners estimate the formant values of a particular speaker's vowel space based on a preceding sentence. When hearing a subsequent speech sound from the same speaker, listeners perceive that target sound relative to a mental frame of reference. This approach focuses much more on a listener's phoneme repertoire and therefore relies on a listener's language background. This framework finds some support in reports such as those by Watkins and Makin (1996), Watkins (1991), Mitterer (2006), and Sjerps et al. (2012), who have shown that a signal-correlated noise precursor induced significantly smaller compensation effects than a speech precursor that had the same LTAS. From a mental frame of reference standpoint, these findings suggest that only a small part of the compensation found with speech sounds is actually due to general auditory processes, while an important contribution is made by a language-specific process such as, possibly, a phonetic frame of reference. In defense of an auditory view, Watkins (1991) and Watkins and Makin (1996) have argued that the lack of compensation effects with noise precursors could be the result of the fact that the noise precursors did not contain spectrotemporal variation. This specific proposal, however, cannot account for the finding that despite similar LTAS relations between precursors and targets, some non-speech precursor signals induce stronger compensation effects than others (Sjerps, Mitterer, & McQueen, 2011a) even though they all had spectrotemporal variation. Sjerps et al. (2011a) therefore suggested that those stimuli that were acoustically (although not necessarily perceptually) more similar to speech induced larger compensation effects. They suggested that language experience could have had an influence on compensation processes. This raises the question, however, through what mechanisms language exposure could potentially impact compensation processes. As will be outlined below, one way in which linguistic exposure could have an influence on compensation processes is through the phonemic inventory of a language.

Perceptual compensation for speaker vocal-tract properties has been shown in at least two languages, Dutch and English (Broadbent & Ladefoged, 1960; Ladefoged & Broadbent, 1957; Mitterer, 2006; van Bergem, Pols, & Beinum, 1988; Watkins, 1991; Watkins & Makin, 1994, 1996). Both of these languages, however, have large vowel inventories, which result in a crowded vowel space with considerable overlap between instances of different vowel categories in the F_1/F_2 vowel space. For listeners of these languages it is beneficial to reduce this overlap by compensating for vocal tract characteristics. In Spanish, a language with a smaller vowel inventory, perceptual confusion among vowel categories is, presumably, less severe than the confusion observed in American English (Cutler, Weber, Smits, & Cooper, 2004). If compensation processes are in part based on an acquired strategy to reduce phoneme confusability, one would expect that listeners of English and Dutch compensate to a larger degree than listeners of Spanish. And this is not the only potential mechanism through which language exposure could influence compensation processes (see below). To address the potential influence of linguistic experience on compensation processes, the current study investigated perceptual adaptation to vocal tract properties in native speakers of Spanish, native speakers of American English and native speakers of Dutch when listening to target vowels embedded in sentences in all three languages. Testing the same participants with materials from different languages, while controlling for their experience with each language, allowed us to probe the contribution of language experience on compensation processes.

In the current study, the strength of compensation effects were tested by means of a target vowel pair that is shared among the three languages and that has no other vowels in between them in any of the stimulus languages. The pair that was considered most suitable was the /o/–/u/ pair, a distinction that lies mainly on F_1 in the three languages. It should be noted that this similarity is true in an abstract phonological sense, and only partly so in its phonetic realization. While F_1 and F_2 are in similar areas in vowel space, both English vowels are diphthongized. For the current study, however, it was the most optimal pair because listeners from all three languages rely on F_1 for the distinction of /o/ and /u/. All listeners heard target vowels on a [sufu]–[sofo] continuum, preceded by Spanish, English and Dutch precursors. These CVCV sequences were chosen because they are non-words in all three languages. Furthermore, the consonants /s/ and /f/ are produced similarly in all three languages. Critically, the precursors were manipulated to have a generally high or a generally low F_1. This design allowed us to examine whether there is an influence of linguistic exposure on the strength of compensation effects, and if so, in what way this influence works. This question thus leads to two hypotheses. The first, most basic hypothesis, is that compensation processes apply to all speech sounds (and to the same extent), irrespective of the language in which the precursor and target are uttered or whether a listener might be familiar with the language. Such a proposal predicts that all listeners, Spanish, English and Dutch, compensate for all precursor sentences (irrespective of the stimulus language). The amount of compensation, in this case, is influenced only by the spectrotemporal characteristics of the signals. Note that the precursor signals for the different languages were necessarily different. The LTAS-based influence of the different materials will therefore inevitably differ to some extent. Our analyses crucially focus on whether listeners from *different* language backgrounds are influenced by the *same* precursors to a different extent.

The second hypothesis states that the combination of a listener's background language and the language of the stimuli plays a role in the strength of compensation processes. This prediction, however, can be borne out in three different ways. The first is that listeners only compensate to the extent that they know a language. Such an interpretation suggests that compensation is related to a listener's subjective impression of listening to their native language. Because language familiarity is a graded dimension we compared perceptual adaptation patterns between native Spanish speakers with low proficiency in English (Spanish) and native Spanish speakers with high proficiency in English (Spanish–English bilinguals). If language familiarity per se can cause differences in the strength of compensation effects, low proficiency speakers of English should compensate less with the English stimuli than high proficiency speakers (and similar effects of familiarity should be found for the other listener–stimulus pairs).

The second way in which language familiarity could influence compensation is through the information value of F_1 in a listener's native language. Native speakers of Spanish, native speakers of English and native speakers of Dutch all identify vowels for an important part on the basis of F_1 and F_2 (and to a much smaller extent F_3, Delattre, Liberman, Cooper, & Gerstman, 1952; Pols, Van der Kamp, & Plomp, 1969). In English and Dutch, however, diphthongization and duration provide additional cues for vowel identity (Adank, van Hout, & Smits, 2004; Strange, 1989). A vowel's exact F_1–F_2 combination is thus likely to be more important for a native speaker of Spanish than a native speaker of English. If the amount of compensation depends on the information value of F_1, it is expected that native speakers of Spanish show the largest amount of compensation.

The third way in which linguistic background could influence compensation effects is through the sound properties of the ambient language. This would suggest that language learners who are exposed to a language with fewer vowel categories, that is, vowel categories which are positioned further apart in the vowel space, will have less of a need to perceptually compensate. Spanish has only five monophthong vowel categories whereas English and Dutch have 11 and 13 respectively. The exact number, though, varies across dialects (Gussenhoven, 1999; Ladefoged, 1999). Note that Spanish also has a number of diphthongs (Aguilar, 1999). However, the important cues for diphthongs are formant movement rather than steady F_1 and F_2 values. According to this version of the language dependency hypothesis, native speakers of Spanish will compensate less than native speakers of Dutch and native speakers of English. This hypothesis, however, rests on the assumption that the vowel categories in Spanish not only lie far apart in vowel space, but also that they have similar within-category variance to those in English and Dutch. If within-category variance in Spanish is much higher than in English and Dutch, the overlap between phoneme categories could still be similar among these languages. In order to examine these properties of the vowel spaces across languages, we collected vowel production data from Spanish, English and Dutch speakers in their native language. We measured the distance between categories and the variance within categories for the five vowels /i/, /e/, /a/, /o/ and /u/.

The goal of the current study was to investigate whether linguistic background has an influence on the strength of perceptual compensation effects, and if so, in what way this influence expresses itself. Native speakers of Spanish, native speakers of English, native speakers of Dutch and Spanish–English bilinguals were tested with materials in all three languages. They categorized target sounds of [sufu] to [sofo] continua which were preceded by sentences that were manipulated to have either a high F_1 contour or a low F_1 contour. This investigation allowed us to test whether the strength of compensation processes is fully determined by signal properties of precursor–target combinations or whether language background has an additional influence.

2. Method

2.1. Participants

Eighteen native listeners of American English (2 male, 16 female), Dutch (4 male, 14 female), Spanish (4 male, 14 female) and eighteen Spanish–English bilinguals (5 male, 13 female) participated in the study. The native English and Spanish–English bilingual participants were recruited and tested in the Phonetics Laboratory at the University of Texas–Austin. The native speakers of Spanish (low English proficiency) were recruited and tested on location at ESL schools in the Austin area and at the Phonetics Laboratory at UT. The native speakers of Dutch were recruited and tested at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands.

Each participant filled a detailed background language questionnaire that was adapted from the LEAP-Q questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007).[1] Additional questions regarding the participants' dialectal background were included. In order to encourage participants to engage in native-language speaking mode, each participant was interviewed prior to the experiment and instructed about the task in their native language by a native speaker. The Spanish–English bilinguals were instructed in Spanish (they mostly started speaking English at arrival but then the experimenter explained that the interview and instructions had to take place in Spanish).

Table 1 provides a summary of a number of measures regarding the language experience for all participants. The results indicate that the four groups of participants were qualitatively different from each other with respect to their dominant language, age of acquisition, the amount of exposure and their proficiency across the three languages. This confirmed our initial assumptions about the differences in the language experience across our listener groups.

2.2. Production

The participants produced the vowels /i/, /e/, /a/, /o/ and /u/ in /sVsV/ and /fVfV/ contexts (where V represents the target vowel). They were asked to produce them in their native language, and the Spanish–English bilinguals were asked to produce them in Spanish. The stimuli were presented to the speakers as a written list on a computer screen (Spanish and English: SISI; FIFI; SOSO; FOFO; SUSU; FUFU; SASA; FAFA; SESE; FEFE, Dutch: SIESIE; FIEFIE; SOSO; FOFO; SOESOE; FOEFOE; SASA; FAFA; SESE; FEFE).[2] Participants read the list of the target non-words three times. Recordings were made using an Audio-technica AT2020 cardoid condenser microphone connected to a TASCAM US-144mkII usb audio interface. The signal was recorded onto a laptop using Adobe Audition software.

The recordings were inspected and a 40 ms portion from the most steady part (for both F_1 and F_2) close to the midpoint of the first vowel in each disyllable was selected. Note that this is not trivial, especially for diphthongized vowels in English. In case no ''most steady part'' could be

[1] The LEAP-Q questionnaire is available at http://comm.soc.northwestern.edu/bilingualism-psycholinguistics/leapq/.

[2] Participants were not provided with a model for the vowels so as to avoid influences on their pronunciation. In case participants used the incorrect vowel, which happened sporadically, they were instructed to repeat the list but now using ''the other vowel''.

Table 1
Average age and language background information for all listeners. 'Dominant language' and 'Acquired first' reflect the percentage of participants that indicated that this language is the most dominant/first acquired language. 'Exposure' is calculated as the average amount of exposure to a language for the participants at this point in their lives. Because responses were unrestricted not all columns may add up to 100 (i.e., participants could indicate that they were also exposed to other languages). 'Proficiency' is calculated as the average over the self-rated proficiency in writing, speaking and understanding speech.

Measure	Listener background			
	Spanish	Bilingual	English	Dutch
Age (yr)	34.6	21.9	20.8	21.6
Dominant language				
Spanish	100	39	0	0
English	0	61	100	0
Dutch	0	0	0	100
Acquired first				
Spanish	100	83	0	0
English	0	17	89	6
Dutch	0	0	0	94
Exposure				
Spanish	59	32	7	0
English	35	67	86	17
Dutch	0	0	0	77
Proficiency				
Spanish	100	86	15	0
English	54	96	100	76
Dutch	0	0	0	100

found, the midpoint was taken. Five formants were estimated between 0 and 5300 Hz using Burg's formant estimation procedure in Praat (Boersma & Weenink, 2009). Measurements for the first two formants were used for further analyses.

2.3. Perception

2.3.1. Recordings and measurements

For each language four carrier sentences were constructed that did not contain the critical sounds /u/ and /o/. These sentences also contained a limited number of sonorant phonemes because these give rise to more artifacts after the resynthesis method (described below). For every language one sentence was selected, based on whether it sounded natural after resynthesis. For Dutch this was an instance of the sentence ''*Die kaas staat niet bij de*'', 'That cheese is not next to the', for Spanish this was ''*a veces se halla*'' 'at times she feels' and for English this was ''*He loves eating fresh*''. The target words were /sofo/ and /sufu/. The precursor sentences were also selected and cut out so that they were of similar length (i.e., 5 or 6 syllables and after splicing out they were all 1.34 s long).

One speaker of each of the three languages was recorded speaking the precursor sentences selected for their native language followed by the target non-words. The Spanish speaker was a speaker of Colombian Spanish, the Dutch speaker a speaker of standard Dutch, and the native English speaker spoke Midwestern Dialect of American English. We obtained a total of 24 instances of both /sufu/ and /sofo/ from each of the three speakers. Measurements of the duration of the target vowels and the trajectories of the first and second formant were taken.

2.3.2. Stimulus manipulation

2.3.2.1. Targets. The vowels of instances of [sofo] (those in both first and second position) were visually inspected and cut out of their fricative context at a zero-crossing. The position was selected as the transition point from the frication into the periodic portion of the vowel or vice versa by looking at the spectrogram and waveform and by listening. The tokens were equalized in duration across the three speakers by cutting out individual periods (first vowel: 147 ms; second vowel: 165 ms). Using Burg's Formant method as implemented in Praat the filter characteristics of the targets were estimated. A source model was estimated with Burg's LPC method, using 80 predictors. Using fewer predictors left remnants of the formants in the signal. Pretesting indicated that such remnants make it harder to induce a shift in the perceptual location of vowel boundaries. One source model of each of the two vowels in [sofo] was selected for each speaker. The onsets and offsets of the source models were ramped in amplitude to correct for differences in the location of zero crossings across tokens.

The filter model was estimated for all vowels by assuming two formants in the range between 0 and 2000 Hz. The trajectory of the F_1 and F_2 were estimated at around 30 points within each vowel (29 for the first vowel and 31 for the second). An average F_1 and F_2 track was calculated over productions of /o/ and /u/ in both the first and second position resulting in a single ambiguous filter model for the first vowel and one for the second vowel (i.e., with values halfway between those in the average [o] and [u], for both formant height and formant bandwidth). This single average was based on a number of repetitions of each of the three speakers. This dynamic filter therefore represented an average of the formant properties over a number of instances of both [o] and [u], and averaged over the three speakers. Fig. 1 displays the average F_1 and F_2 tracks for the initial target vowel. These contours were used to resynthesize the target vowels for the three speakers, so that their first and second formant tracks would be identical. This procedure was followed so that the final manipulated target words had the same filter properties across the three different speakers. This assured that listeners could only rely on the same cues in the targets across the stimulus languages. In steps of 10 Hz, the height of only the first formant of the filter model was increased over a range of 100 Hz and decreased over a range of 100 Hz across the whole vowel to create the new formant models for the continuum from [u] to [o] (now only distinguished by F_1). The resulting formant models were then reapplied to the source models of a single instance of the first and second vowel of [sofo] for every speaker. So for each speaker the

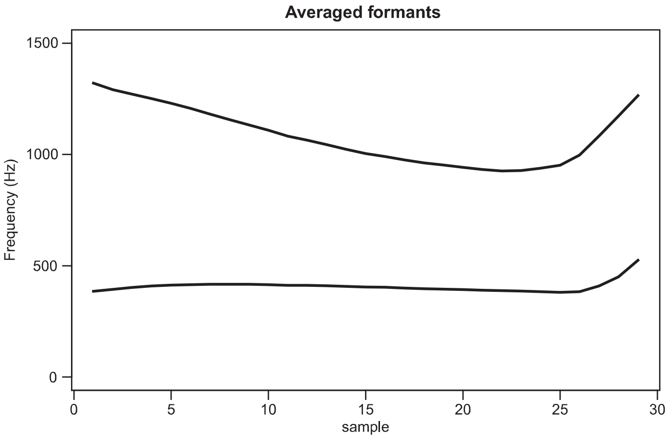

Fig. 1. Averaged formant contours for the first and second formant of vowels that were recorded in a sVfV context (only the contours of the first vowel are displayed). See text for further detail.

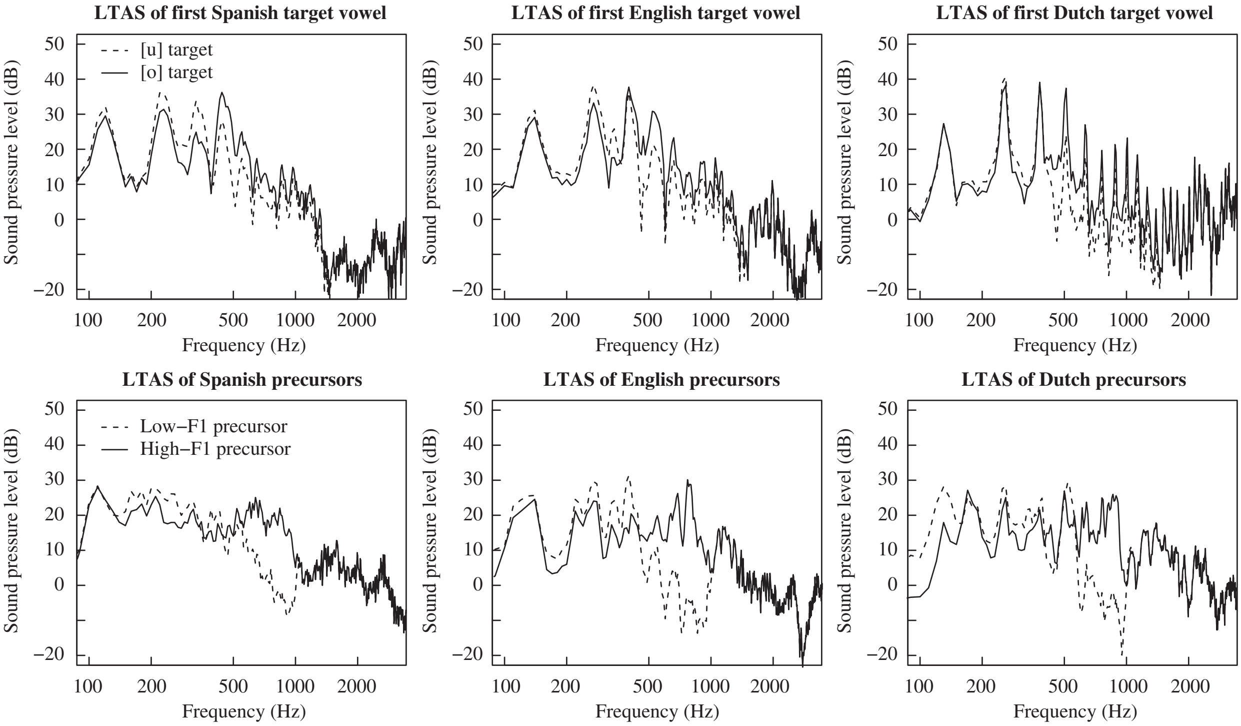

Fig. 2. LTAS plots for the stimuli. Top panels: LTAS for the endpoint target vowels ([u]=dotted line, [o]=solid line) for, from left-to-right, the Spanish, English and the Dutch materials. Bottom panels: LTAS for the precursor sentences (Low F_1=dotted line, High F_1=solid line) for, from left-to-right, the Spanish, English and the Dutch materials. The *x*-axes are logarithmic.

average filter properties were used but they were combined with a speaker-specific source model. The resulting signals were low-pass filtered between 0 and 1500 Hz (with the standard filter function in Praat that filters in the frequency domain with a smoothing of 100 Hz). These sounds were modified to have the same amplitude trajectory and overall amplitude as a low-pass filtered instance of the original vowel that was used for the source model. These sounds were then added to the high pass filtered parts (1500–6000 Hz) of the original vowels that had been used to create the source models (through summation of the signals across time). Vowels created in this way were spliced into the consonantal contexts of the particular speaker (the consonants were also equalized in duration across the speakers). All created items were equalized in amplitude. This yielded a total of 21 tokens of the [sofo] to [sufu] continuum per language/speaker.

We determined the 50% categorization-crossover points of the two vowels through pilot testing. Five listeners categorized the targets of each of the three languages in a forced choice task from /sofo/, /sufu/, /sofu/ and /sufo/ response options. Based on this pretest, a 7-step range of vowels that covered an F_1 range of 120 Hz was selected. Each step was 20 Hz, ranging from, on average, 337 to 457 Hz for the first vowel and from 359 to 479 Hz for the second vowel. The top panels of Fig. 2 display the LTAS of the endpoint vowels (those in initial position). The LTAS were calculated in Praat with a 10 Hz bin-width. Note that the *x*-axis is logarithmic. The left-hand panel of Fig. 3 displays the differences between the LTAS of the endpoint target vowels ([o]–[u]) It can be observed that the differences are reasonably similar across the stimulus languages, which is expected since the same formant filter model was used for the resynthesis of the vowels.

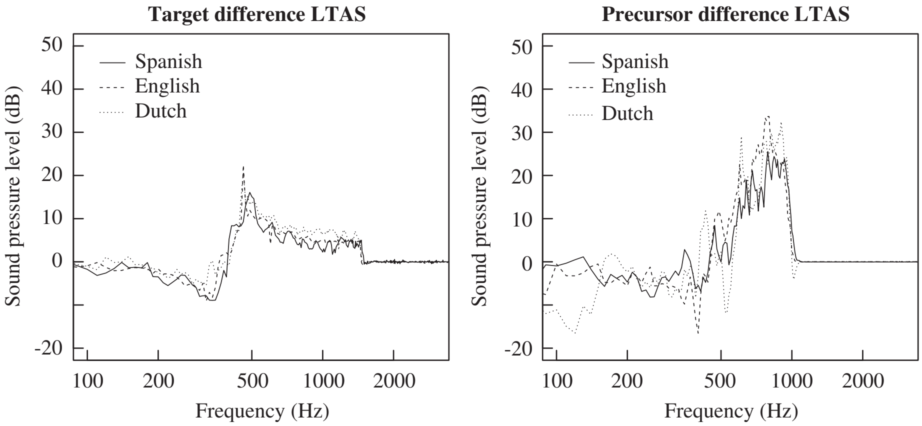

Fig. 3. Difference LTAS lines. Left panel: difference LTAS for the endpoint targets for the three different language materials (Spanish = solid; English = dashed, Dutch = dotted). Right panel: difference LTAS for the precursors in the different stimulus languages (Spanish = solid; English = dashed, Dutch = dotted). The *x*-axes are logarithmic.

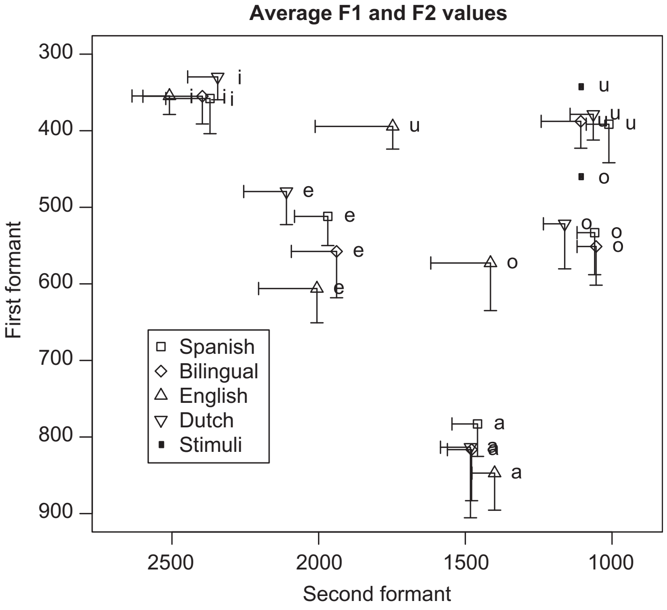

Fig. 4. Average formant values plotted in F_1–F_2 space, for four groups of speakers (Spanish, bilinguals, English and Dutch), each with 18 speakers. The target endpoint values of the first vowel for the range of stimuli used in the listening experiment are also displayed. Whiskers represent one standard deviation and are based on the same data as reported in Table 2.

2.3.2.2. Precursors. The source and filter models of the precursors were estimated with Burg's method in Praat, using the same parameters as for the targets. The F_1 track of the filter model was increased or decreased by 200 Hz (values are based on Watkins & Makin 1994) for each of the three speakers to create two new versions, one with a low and one with a high F_1 contour. One formant was estimated between 0 and 1000 Hz and the original signal was used for frequencies above 1000 Hz. This adaptation resulted in more natural sounding precursors. As with the targets, the precursor sound files were equalized for overall intensity and duration. Finally, the precursor sentences and targets were concatenated with a 500 ms silent interval between them. The relatively long precursor–target interval was used to minimize peripheral auditory influences. The bottom panels of Fig. 2 display the LTAS of the two contexts in each of the languages. The right-hand panel of Fig. 3 displays the differences between the LTAS of the two context versions in each of the three languages high F_1–low F_1. Although the differences between the different language materials are somewhat larger than for the targets, the differences are still reasonably small. This indicates that, based purely on the LTAS, the effect sizes should be similar across the stimulus sets.

2.3.3. Procedure

After arriving at the testing site, the participants first filled out a language background questionnaire (LEAP-Q, Marian et al., 2007). Following that, they were recorded reading the list of non-words. Finally, they participated in the listening test. The different language-materials were presented in each of the six different possible orders to different participants following a Latin-square design. Each different ordering was presented to three participants of a particular language background. Listeners received written instructions about the task in their native language (a native speaker of their language was present to answer any additional questions). The participants were asked to categorize the last words of a stimulus as either /sufu/ or /sofo/. Listeners responded using the two buttons located below the touchpad on the laptop. The two options "sufu" and "sofo" were always displayed on the computer screen. A practice session preceded the test in order to familiarize the listeners with the task. Each of the 7 steps of the continuum was presented in both the low- and high-F_1 sentence conditions, randomly intermixed. Such a block was repeated 8 times per exposure language, resulting in 336 trials (2 precursor conditions × 7 steps × 8 repetitions × 3 languages). Stimuli were presented using Presentation software (Version 11.3, Neurobehavioural Systems Inc.). The listening task took roughly 30 min per participant.

3. Results

3.1. Production data

Fig. 4 displays the average $F_1 \times F_2$ values for each vowel for each group of speakers. Each symbol represents an average over approximately 108 vowel tokens (18 speakers × 2 non-words × ∼3 repetitions; the number of repetitions is approximate as some instances, such as clear mispronunciations, were discarded). Participants were instructed to produce vowels in their native language (bilinguals were asked to say them in Spanish) in /sVsV/ and /fVfV/ context, where V represents the vowel. Measures are based on a 40 ms window in the steady part of the first vowel. The two filled squares represent the two endpoint stimuli used in the perception experiment. Note that for the stimuli the formant values appear relatively low. This is due to the fact that the speakers for the stimuli were male, whereas most participants were female speakers, who tend to have higher formant values in general (Hillenbrand, Getty, Clark, & Wheeler, 1995). The figure shows that the five vowel categories occupy similar positions in the vowel space in all three languages. The exceptions are the English /o/ and /u/ which have a relatively high F_2 value, that is, they are fronted. This is a known phenomenon, especially in the Southern varieties of American English (AE) (Clopper, Pisoni, & de Jong, 2005). Table 2 reports the between-participant variance of F_1 and F_2 in productions of /i/, /e/, /a/, /o/ and /u/ for speakers of the different language backgrounds. The results showed that the productions of the speakers from the different groups displayed similar amounts of variance. These results provide evidence that the overlap across various vowel categories in the F_1–F_2 space may be reduced in Spanish (because the five plotted vowels constitute the full Spanish monophthong inventory) compared to Dutch and English (which both have additional vowels lying between the plotted vowels). Note that the different groups have different numbers of male participants contributing to the measures (American-English: 2; Dutch: 4; Spanish: 4; bilinguals: 5 males). If anything, males (having lower formant values overall) would be expected to show less variability. The number of males per group does not seem to have a strong influence on the data (e.g., the bilinguals do not show lower variances in the formant values). We calculated the standard deviations for each gender group separately and then calculated a weighted mean across genders (see formula at the bottom of Table 2).

3.2. Perception

3.2.1. Overall analysis

The categorization data were logit-transformed (Dixon, 2008). These data were then analyzed using repeated-measures ANOVAs. An omnibus analysis included the within-subject factors Language (defined by the stimulus materials, with the levels Spanish, Dutch, English), Context (defined by the precursors, which had either a high or a low F_1) and Step (defined by the stimulus continua consisting of seven steps from [sufu] to [sofo]), and the between-subject factor Listener (defined by participant backgrounds, with the levels Spanish, bilingual, Dutch, English). In cases of significant departure of sphericity a Greenhouse–Geisser correction was applied.

Fig. 5 displays the categorization results split out over the different stimulus materials and background language groups. Fig. 6 displays the overall Context effects for those groups, averaged across the continuum. The bars represent the numerical size of the compensation effect (p(/sofo/) in low F_1-p(/sofo/) in high F_1). Overall, a strong effect of Context was found which indicated that listeners more often gave /sofo/ responses after a precursor with a low F_1 than a precursor with a high F_1: $F(1, 68)=106.9$, $p<0.001$, $\eta^2_p=0.611$. This can be observed in Fig. 5 as

Table 2
Between-speaker variation (standard deviations) of the F_1 and F_2 in productions of /i/, /e/, /a/, /o/ and /u/ for speakers of four different language backgrounds. Participants uttered /sVsV/ and /fVfV/ non-word sequences (where "V" represents the vowel). Measures are based on the first vowel. Measures were taken of the F_1 and F_2 values in those productions. Reported data represent the between-speaker variance, measured as the standard deviation over 18 participants per group. Values per participant were based on an average over ∼3 repetitions of each non-word.

Vowel	Listener background			
	Spanish	Bilingual	English	Dutch
Measure				
F_1				
/i/	49.9	38.9	23.7	29.9
/e/	38.6	62.7	45.4	46.6
/a/	46.6	67.1	48.6	92.5
/o/	59.1	50.7	61.1	60.4
/u/	54.3	34.8	29.1	33.5
Average	49.7	50.8	41.6	52.6
F_2				
/i/	159.0	222.3	125.3	104.4
/e/	113.8	166.3	198.1	145.3
/a/	97.2	80.3	77.2	102.7
/o/	60.6	66.8	200.4	76.3
/u/	77.2	132.0	260.9	79.7
Average	101.6	133.5	172.4	101.7

Note: For listeners of the same language background, standard deviations were calculated separately for speaker of different gender. These were then combined by calculating a weighted average (using the following formula: $\mathrm{sd}_{FM}=\sqrt{(((n_F-1)\mathrm{sd}_F^2+(n_m-1)\mathrm{sd}_m^2)/(n_f+n_m-2))}$; note that this approach assumes equal variances among males and females).

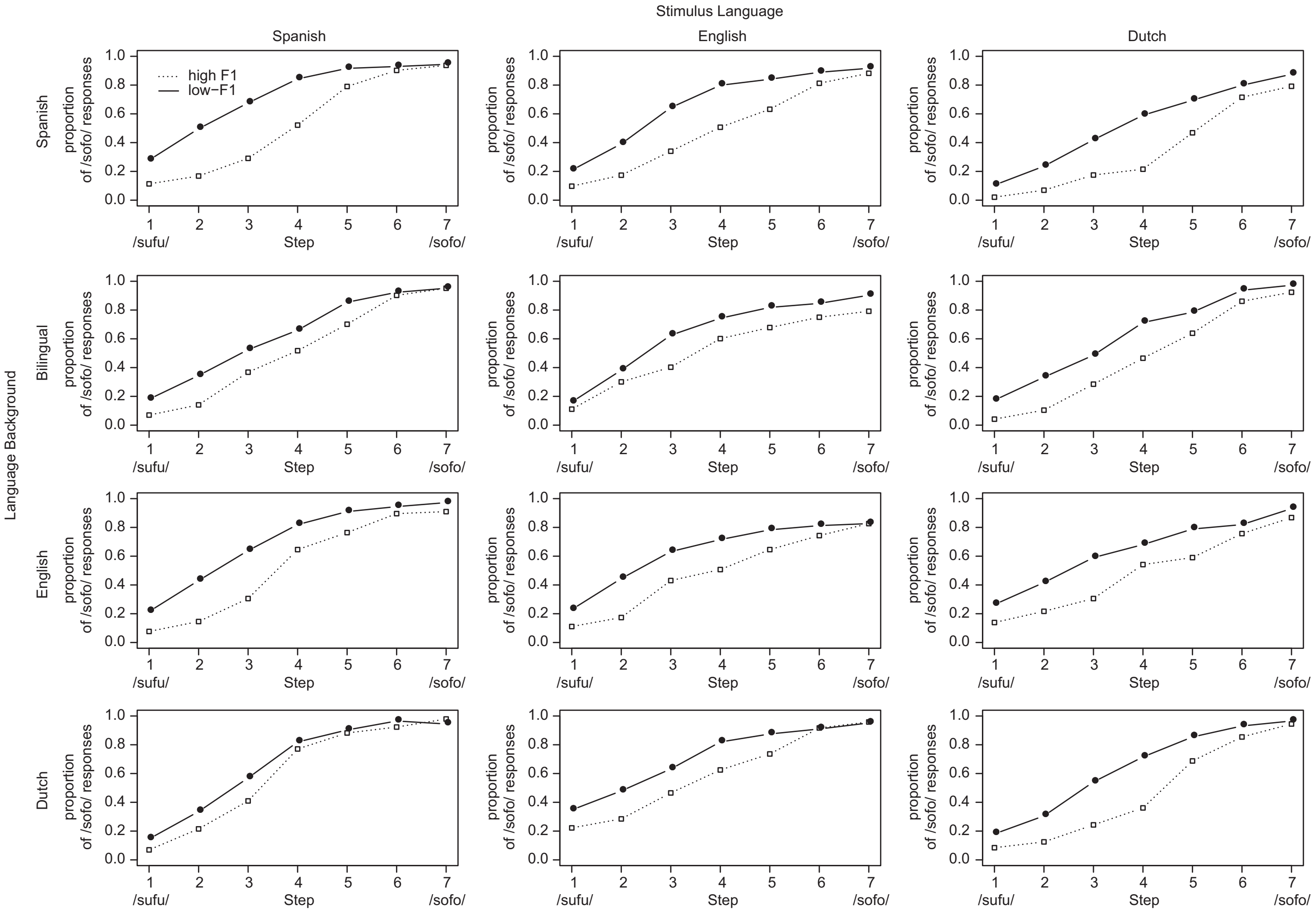

Fig. 5. Categorization curves representing the probability of a /sofo/ response to targets of a 7 step [sufu] to [sofo] continuum. Targets were presented in the context of a precursor sentence that was manipulated to have either a generally low F_1 (solid line) or a generally high F_1 (dotted line). From left-to-right columns display the stimuli spoken in: Spanish, English and Dutch. From top-to-bottom, rows display data from the listeners that had a Spanish, Spanish–English bilingual, English and a Dutch language background.

the separation between the solid and the dotted lines. This is a reflection of the spectrally contrastive compensation effect because more high F_1 targets (/sofo/) were chosen after a low F_1 precursor and vice versa. Furthermore, participants reliably categorized the continuum, with more /sofo/ responses to a target with a high F_1, and vice versa: $F(1.97, 408)=687.4$, $p<0.001$, $\eta_p^2=0.910$. The analyses revealed that the different materials led to different overall numbers of /o/ responses: $F(2, 136)=6.99$, $p=0.001$, $\eta_p^2=0.093$. This can be observed from Fig. 7, which displays the average categorization curves for the three sets of materials (across all listener groups and context conditions). The analyses also established that the shapes of the categorization curves differed across the language materials, expressed as an interaction between Language and Step: $F(5.66, 816)=8.2$, $p<0.001$, $\eta_p^2=0.108$. Moreover, the effect of Context differed across the continuum, as expressed by an interaction between Context and Step: ($F(4.109, 408)=41.531$, $p<0.001$, $\eta_p^2=0.379$). Furthermore, three just non-significant interactions were observed. These suggest that the effect of Context was expressed at different parts of the continuum for different materials, and for the different groups of participants. These effects were reflected as a marginal three-way interaction between Context by Step by Listener: $F(18, 408)=1.55$, $p=0.069$, $\eta_p^2=0.064$; as a marginal interaction between Language by Context by Step: $F(8.55, 816)=1.73$, $p=0.083$, $\eta_p^2=0.025$; and as a marginal interaction between Language by Context by Step by Listener: $F(36, 816)$, $p=0.084$, $\eta_p^2=0.056$.

Importantly, non-significant effects were observed for the critical interactions between Context by Listener: $F(3, 68)=1.19$, $p=0.322$, $\eta_p^2=0.050$; the interaction between Context and Language: $F(2, 136)=1.36$, $p=0.26$, $\eta_p^2=0.20$; and the three-way interaction between Language by Context by Listener: $F(6,136)=1.216$, $p=0.302$, $\eta_p^2=0.051$.

4. General discussion

The current paper investigated the effect of language background on perceptual accommodation to speaker vocal-tract variation. Listeners categorized vowel targets on a [sufu] to [sofo] continuum. These targets were preceded by precursor sentences that had either a high or a low F_1 (following Ladefoged & Broadbent, 1957). Listeners from four language backgrounds participated in this study. These were native speakers of Spanish, Spanish–English bilinguals, native speakers of English and native speakers of Dutch. All participants listened to stimuli in Spanish, English and Dutch. This design allowed us to investigate language-dependent influences on perceptual compensation for speaker vocal tract characteristics while controlling the physical stimulus characteristics.

The results showed that listeners from all language backgrounds compensated for the vocal tract properties of a speaker in a precursor sentence. While previous reports have documented compensation effects in Dutch (Mitterer, 2006; Sjerps et al., 2011a; van Bergem et al., 1988)

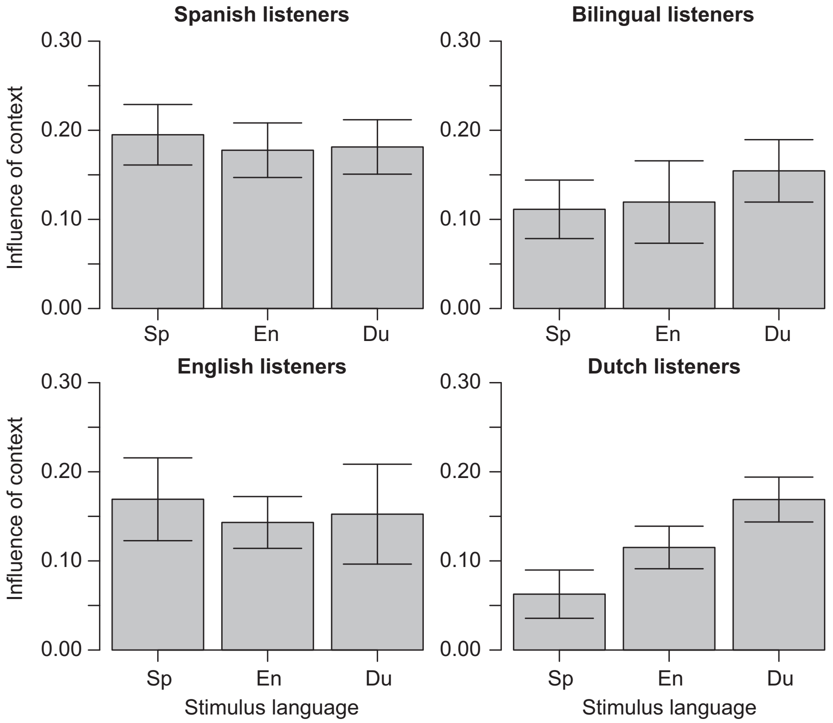

Fig. 6. Bar-graphs representing the average difference between the high-F_1 and the low-F_1 categorization curves. Each panel displays the effects for the three stimulus languages Spanish, English and Dutch. From left-to-right and from top-to-bottom the panels display the data for listeners that had a Spanish, bilingual, English or a Dutch language background. Error bars reflect standard errors of the mean.

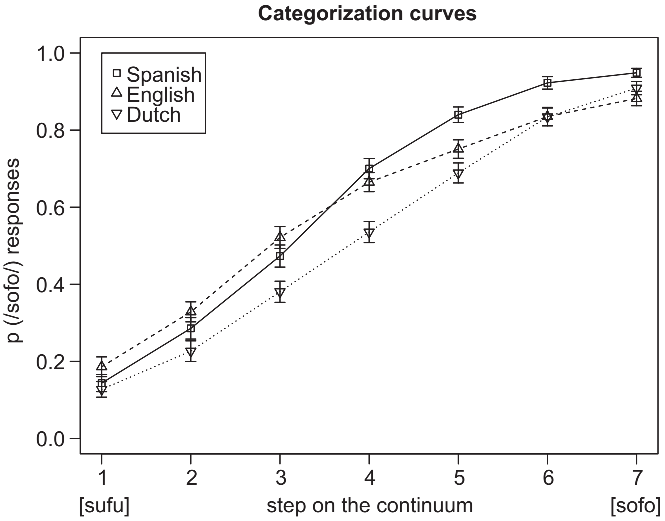

Fig. 7. Proportion of /sofo/ responses for the 7 steps on the stimulus continua for the materials spoken in Spanish (squares, solid lines), English (upward facing triangles, dashed line) and Dutch (downward facing triangles, dotted line). Error bars reflect standard errors of the mean.

and English (Ladefoged & Broadbent, 1957; Watkins, 1991; Watkins & Makin, 1996), the current study extends this effect to a Romance language with different vowel properties. Compensation for vocal tract characteristics thus seems to be a general property of listening to speech. Furthermore, listeners compensated for precursors when they listened to the materials in their own language as well as when they listened to the materials in their second language and in an unfamiliar language. This result emphasizes the generality of the mechanisms that cause compensation for vocal tract characteristics.

One of the most critical findings of this study, however, is that the perceptual impact of a precursor sentence on subsequent targets was similar in size across participants from different language backgrounds. Previously, differences in the amount of compensation have been shown to occur across different types of manipulated speech sounds (Sjerps et al., 2011a). The current study, however, shows that compensation effects are quite similar for speech stimuli, irrespective of a listener's familiarity with that language. At the outset of this research we had formulated two hypotheses. The most basic hypothesis was that compensation is completely independent of specific influences of language background and stimulus languages. The other hypothesis postulated that linguistic experience could have an influence on the size of compensation effects. A language specific influence could express itself in three different ways: through overall familiarity to a stimulus language; through the informativeness of F_1 in a listener's native language; or as an adaptive mechanism to deal with vowel overlap in the native language.

With respect to the latter, the production results indeed revealed that between-speaker variance was similar across the four groups of participants. Since English and Dutch have a number of additional vowels in between the vowels that are shared with Spanish it is possible that English and Dutch have more category overlap than Spanish. The categorization results, however, revealed that listeners' language backgrounds did not influence the strength of compensation effects. These data therefore do not lend support to any of the tested versions of a hypothesis that relies on language background as an explanatory factor for the strength of context effects.

The categorization results in the current study suggest that compensation effects with speech materials are for an important part determined by acoustic properties of the stimuli as per our first hypothesis. This conclusion thereby aligns with a number of findings on compensation processes involved in both the perception of vowels and consonants (Laing et al., 2012; Sjerps et al., 2011a; Watkins & Makin, 1996). For instance, Laing et al. (2012) recently report a contrastive shift in categorization of tokens on a [ɡa]–[da] continuum (a distinction mainly determined by F_3) when these were preceded by a sentence that had been synthesized to have either a heightened or a lowered F_3. Importantly, a similar shift was observed when these sentences were replaced by a sequence of sinusoidal tones (with, similarly, a high and a low spectral version) with a distribution in the same frequency region as the F_3 of the target distinction. In the study by Laing et al. (2012), however, the speech and non-speech precursors naturally consisted of different signals and only the position of spectral prominences was matched between the signals, not the size of the spectral difference between the two context conditions. This aspect precludes a direct comparison between differences in the size of the effects across speech and non-speech materials. The approach taken here, relying on the use of stimuli from languages that are unintelligible to some of the participants, is a useful way to specifically address the aspects of intelligibility and previous exposure that are only a subset of the dimensions captured by speech versus non-speech comparisons. This allowed for a more straightforward comparison of effect sizes across conditions. Nonetheless, the current findings are in line with Laing et al. (2012) and point towards relatively general auditory processes as the main explanatory factor in contrast effects in speech perception.

5. Conclusion

Perceptual compensation processes aid listeners in dealing with across-talker variation in speech. Such compensation mechanisms may increase the effective transfer of information in perception (Kiefte & Kluender, 2008; Kluender et al., 2003; Kluender & Kiefte, 2006). Evidence is accumulating that these processes are based for an important part on general auditory processes (Sjerps et al., 2011a; Sjerps, Mitterer, & McQueen, 2011b; Stilp et al., 2010; Watkins, 1991; Watkins & Makin, 1994, 1996). Adding to this growing body of research, the current study shows that compensation effects can be found with the same materials across three different languages and for listeners who are listening to a second language or an unfamiliar language. The size of the compensation effect is mainly dependent on the acoustic properties of speech sounds, not on a listener's familiarity to the stimulus language.

Acknowledgments

This research was supported by a visiting scholar stipend awarded to the first author by the University of Austin Texas. We would like to thank Anais Callejon, Rob Corona and Randall Rouse for their assistance in testing participants, and Daniel Olson, Omar Vargas and Jelmer Wolterink for recording the stimuli. We would also like to thank Mike Smith from UT ESL Services and Suzanne Baker, Career Counsellor at Capital IDEA, for their support in recruiting native Spanish participants. We are grateful to James McQueen and Holger Mitterer for their suggestions on the design of the experiments and comments on earlier versions of the manuscript. Finally, we thank the anonymous reviewers and the editor for their feedback.

References

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America*, *116*(3), 1729–1738.

Aguilar, L. (1999). Hiatus and diphthong: Acoustic cues and speech situation differences. *Speech Communication*, *28*(1), 57–74.

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer* (Version 5.1).

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgements and adaptation level. *Proceedings of the Royal Society of London Series B—Biological Sciences*, *151*(944), 384–399.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America*, *118*(3), 1661–1676.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, *116*, 3668.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*, 195–210.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456.

Gussenhoven, C. (1999). *Dutch handbook of the International Phonetic Association* (pp. 74–77). Cambridge: Cambridge University Press.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312.

Holt, L. L. (2006). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America*, *119*(6), 4016–4026.

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*(1–2), 156–169.

Johnson, K. (2005). Speaker normalization in speech perception. In: D. B. Pisoni, & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford: Blackwell.

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. *Journal of the Acoustical Society of America*, *123*(1), 366–376.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*(1), 59–69.

Kluender, K. R., & Kiefte, M. J. (2006). Speech perception within a biologically realistic information-theoretic framework. In: M. A. Gernsbacher, & M. Traxler (Eds.), *Handbook of psycholinguistics* (2nd ed.). London: Elsevier.

Ladefoged, P. (1999). *American English handbook of the International Phonetic Association* (pp. 41–44). Cambridge: Cambridge University Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98–104.

Laing, E., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology*, *3*, 203.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *Journal of the Acoustical Society of America*, *113*(1), 53–56.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech Language and Hearing Research*, *50*(4), 940–967.

Mitterer, H. (2006). Is vowel normalization independent of lexical processing?. *Phonetica*, *63*(4), 209–229.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088–2113.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pols, L. C. W., Van der Kamp, L. J. T., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America*, *46*, 458–467.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011a). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics*, *73*, 1195–1215.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia*, *49*, 3831–3846.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2012). Hemispheric differences in the effects of context on vowel perception. *Brain and Language*, *120*, 401–405.

Stilp, C. E., Alexander, J. M., Kiefte, M. J., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics*, *72*, 470–480.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America*, *85*(5), 2135–2153.

Summerfield, Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. *Perception & Psychophysics*, *35*(3), 203–213.

van Bergem, D. R., Pols, L. C. W., & Beinum, F. J. K.-v. (1988). Perceptual normalization of the vowels of a man and a child in various contexts. *Speech Communication*, *7*(1), 1–20.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *90*(6), 2942–2955.

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *96*(3), 1263–1282.

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *99*(6), 3749–3757.

Wilson, J. P. (1970). An auditory after-image. In: R. Plomp, & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (pp. 303–318). Leiden: Sijthoff.