“Excuse meeee!!”: (Mis)coordination of lexical and paralinguistic prosody in L2 hyperarticulation

Yuki Asano, Michele Gubian

PII: S0167-6393(17)30201-7
Reference: SPECOM 2518

To appear in: Speech Communication

Received date: 30 May 2017
Revised date: 6 November 2017
Accepted date: 18 December 2017


This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
“Excuse meee!!”: (Mis)coordination of lexical and paralinguistic prosody in L2 hyperarticulation

Yuki Asano and Michele Gubian

Corresponding author:
Yuki Asano
University of Tübingen
Wilhelmstraße 50
72076 Tübingen
Germany
Yuki.Asano@uni-tuebingen.de

Michele Gubian
School of Experimental Psychology
University of Bristol, UK
Office 3D3
The Priory Road Complex,
Priory Road, Clifton BS8 1TU
e-mail: mm14722@bristol.ac.uk
Abstract

The present study investigates how first language (L1) and second language (L2) speakers coordinate lexical and paralinguistic prosody (lexically determined pitch accent and utterance-level intonation) in hyperarticulated speech. Modifications of $F_0$ and segmental durations in repeated utterances were analyzed jointly using a novel method, Functional Principal Component Analysis (FPCA). By testing Japanese and German participants as L1 and as L2 speakers of the respective languages, the task elicited a situation in which L2 speakers had to handle $F_0$ and segmental durations when these prosodic properties mainly carry a paralinguistic function in their L1, but carry a lexical one in their L2 or vice versa. Results show language-dependent and -independent modifications in hyperarticulated speech in both speaker groups. The language-dependent modifications were negatively transferred to the respective L2. The restriction of Japanese lexical pitch accent prevented Japanese participants from phonologically modifying $F_0$ contours in L1 and L2, and German L2 speakers applied the German pitch accent rules in L2 Japanese utterances. Lexical rules outweighed paralinguistic rules in modifying prosody. At the same time, both German and Japanese speakers hyperarticulated the utterances with utterance-final lengthening, larger pitch range and higher peak. Furthermore, we argue that there may be different dominance orders of prosodic properties in Japanese and German according to which prosodic property was more likely to be modified in hyperarticulated speech. L2 speakers were more prone to change a more dominant prosodic property in their L1 to convey a paralinguistic meaning, i.e., segmental durations for Japanese and $F_0$ for German. The findings are particularly noteworthy as the uttered words were highly frequent words of which L2 learners should have had sufficient L2 input before.

Key words

lexical vs. paralinguistic prosody, L1 & L2 hyperarticulation, Functional Principal Component Analysis, Functional Data Analysis
1 Introduction

Imagine being in a crowded and noisy bar where you are trying to call a waiter to order a drink. The waiter does not hear you calling and you have to try to call him again. In such a situation, a speaker often reinforces and hyperarticulates an utterance in order to pursue a response (Stivers and Rossano, 2010a,b), by changing the syntactic structures or by changing prosodic properties\(^1\) (Stent et al., 2008). The current study focuses on the latter, the prosodic changes, especially focusing on intonation and segmental length which are measured by \(F_0\) and segmental durations, respectively. Now imagine being in a crowded and noisy bar in a foreign country. Should you try to catch the waiter’s attention and pursue response, like in the situation above, you would be willing to reinforce the utterance by changing L2 prosody. Difficulties may occur in such a situation when the uses of L2 and L1 prosody are different. In some languages, for instance in Japanese, prosodic properties such as \(F_0\) and segmental length are primarily used to convey lexical information and distinguish words (Sekiguchi and Nakajima, 1999); while in other languages, for instance in German or English, these prosodic properties primarily serve at the paralinguistic level to convey emotion or attitudinal states of a speaker (Baumann and Grice, 2006; Gibbon, 1998; Liscombe, 2007) or at the post-lexical level to signal syntactic or pragmatic information such as topic vs. focus, sentence mode (question vs. statement) (e.g., Braun, 2006; Féry, 1993). What matters is that lexical and paralinguistic prosodic properties have to be coordinated without competing with each other. The question investigated in the present study is how a speaker reinforces an utterance in an L2 when the same prosodic features in the L1 and L2 are primarily used at different linguistic levels (lexical vs. paralinguistic levels), such as in the case reported between Japanese and German.

The difficulties in acquiring L2 prosody are well known since almost every adult L2 learner retains an immediately identifiable foreign accent, and such a foreign accent has also been observed in otherwise highly proficient L2 speakers who mastered the grammatical system very well (Altmann and Kabak, 2011; Els and Bot, 1987; Gut, 2003; Jilka et al., 2007; Mennen, 2007; Oyama, 1976). One of the major factors causing a foreign accent is negative language transfer from one’s L1 to L2 (Gass and Selinker, 1994), which has been an important issue in L2 acquisition research. Negative transfers are not limited to the acquisition of novel sounds at the segmental domain (e.g., the distinction between /r/ and /l/ by Japanese learners of German) but also extend to the prosodic domain (Trouvain and Gut, 2007, for an overview). However, the number of studies investigating the latter domain is still smaller compared to the former (Asano, 2016). As a consequence, the theories of phonological acquisition consider phenomena mostly at the segmental level, but rarely at the prosodic level, and mostly focus on L2 perception (Best, 1995; Best et al., 2001; Best and Tyler, 2007; Flege et al., 1995; Flege, 1999; Flege et al., 2002; Kuhl, 1991; Kuhl and Iverson, 1995) rather than on production (with some excep-

\(^1\)The term prosody broadly refers to the phonological organization of individual sounds (i.e., segments) into higher-level constituents, which is cued by variation of \(F_0\), duration, amplitude and segment quality (Shattuck-Hufnagel and Turk, 1996; Ueyama, 2000). Since phonetic correlates of stress and intonation often manifest at the segmental level, we will use the more general term “prosody” for these phenomena instead of using the division of segmentals vs. suprasegmentals.
tions such as De Bot, 1992; Kormos, 2006, 2012). Therefore, the incorporation of L2 prosody into models of speech processing and acquisition is still awaited.

In this work, we will investigate Japanese and German\(^2\) as an example of two languages whose prosodic systems contrastively differ from each other. Regarding the differences in the use of \(F_0\), in Japanese, the meaning of a word can change depending on the position of a lexically specified pitch fall, called a pitch accent (Kubozono, 2011; Vance, 1987), e.g., \(\text{hashi-ga} = \text{chopsticks}_{\text{NOM}}, \text{hashi-ga} = \text{bridge}_{\text{NOM}}, \text{hashi-ga} = \text{edge}_{\text{NOM}}\), the grave accent indicates the position of the pitch fall, if specified. In other words, the presence or absence of a pitch accent is an inherent property of a word. In German, \(F_0\) is one of the phonetic constituents besides duration and intensity that build a metrical word stress (Wiese, 2000). In other words, a stressed syllable receives a pitch movement. Crucially, however, the direction of a pitch movement is not lexically fixed, i.e., both rising and falling pitch accents are allowed depending on the post-lexical or paralinguistic information that a speaker intends to convey (Baumann and Grice, 2006; Braun, 2006; Féry, 1993; Gibbon, 1998; Grice et al., 2005) (see Figure 1).

![Figure 1: Example contours of the word Entschuldigung, one with a falling pitch accent on the stressed syllable (schul) (left), the other with a rising pitch accent (right).](image)

With respect to a phonological and phonetic form of a pitch accent, in Japanese there is only one phonological type of a pitch accent that is phonetically realized as a sharp fall from a high level occurring near the end of the accented mora to a low level in the following mora (Gussenhoven, 2004; Vance, 1987; Venditti, 2000). This phonetic form does not exist in German as a falling pitch accent. In German, the phonological inventory of pitch accents is richer (six basic pitch accent types) (Baumann et al., 2001; Grice et al., 1996) and each accent shape is associated with a specific pragmatic meaning that contributes to the overall meaning of the utterance, as in English (Pierrehumbert and Hirschberg, 1990). Finally, the phonetic correlates of a pitch accent also differ in the two languages. Japanese pitch accents are characterized only in changing of \(F_0\), while German pitch accents relate to the phonetic changes in \(F_0\), duration and intensity (Wiese, 2000).

Japanese pitch accent does not trigger a longer duration or a higher intensity, because these cues contribute to the perception of acoustic length, which can change the meaning of a word. Japanese is a mora-timed language (Cutler and Otake, 1999; 2000).

\(^2\)In this paper, “Japanese” refers to standard Tokyo Japanese and “German” to standard German unless otherwise stated.
Pike, 1946): i.e., a syllable consisting of a vowel or a vowel and consonant takes up one timing unit (mora) and all morae have approximately the same duration. Japanese distinguishes between long and short vowels as well as long and short consonants (e.g., *kite* = *come*, *ki:te* = *listen*, *ki:te* = *cut*, all in the imperative form - the colon indicates a long segment) (Kubozono, 1993, 1999). On the other hand, German is a stress-timed language in which stressed syllables occur at regular intervals. In German, only vowel length is contrastive for */a/ and */e/ (Wiese, 2000), while consonant length is not contrastive\(^3\).

The different uses of the same prosodic properties in Japanese and German may become hurdles in L2 acquisition. If negative transfer occurs from German to Japanese, then German learners of Japanese may risk communication problems since words may not be identifiable. In case of transfer from Japanese to German, Japanese learners of German may deliver undesirable attitudes without being aware of it. For instance, a falling pitch accent, described as H\(^+\)+L L-% in the Japanese ToBI system (Venditti, 1997), is the lexical default of a Japanese pitch accent without conveying a certain emotional status, while this realization in German is one of many phonological varieties and conveys a certain emotional meaning such as frustration or seriousness (Baumann et al., 2001; Grice and Baumann, 2002). If Japanese learners of German are able to produce only this default Japanese phonological form in German L2, regardless of their emotions or situations, they may sometimes risk conveying undesired paralinguistic information.

Our study aims at testing how L2 learners whose L1 and L2 exhibit prosodic differences, like those found between Japanese and German, manage to coordinate lexical and paralinguistic prosody to produce an L2 utterance in an appropriate way. To achieve this end, an experiment was conducted to elicit a situation in which lexical and paralinguistic properties of these languages compete against each other. In a simulated attention-seeking situation, speakers were asked to repeat the same utterance (*Excuse me* in Japanese and German, to call a waiter) and were instructed to imagine becoming frustrated. We analyzed how L1 and L2 speakers of Japanese and German changed their $F_0$ contours and segmental durations to convey this paralinguistic (emotional) information. The crucial point is that the Japanese utterance *sumimasen*, “excuse me”, was bound to a lexically fixed falling pitch accent, while the German *Entschuldigung*, “excuse me”, was not, i.e., both falling and rising pitch accent were possible depending on the paralinguistic information that a speaker intended to convey.

The investigation carried out in this study presents several elements of novelty. First, we investigated bi-directional language transfer as defined in Ueyama (2000), namely language A as L1 to language B as L2 and language B as L1 to language A as L2\(^4\). Most previous studies have examined only one direction of language transfer (i.e., either transfer from language A to language B or visa versa), while only a few studies of L2 speech production have examined the bi-directional prosodic transfer within one single study (e.g., Trouvain et al., 2016; Ueyama, 2000; Wester et al.,

\(^3\)Consonant length contrasts may be lexically used in Swiss German and in some German dialects such as in Alemannic German (Kraehenmann, 2001; Kraehenmann and Lahiri, 2008).

\(^4\)Note that the term “bi-directional language transfer” in this paper is not used to refer to language transfer from L1 to L2 and L2 to L1, as Mennen (2004) has reported.
2014). Bi-directional analysis enables us to examine cross-linguistic similarity and the symmetry in transferring, thus is more informative in investigating prosodic transfer than the analysis of only one direction. Second, we examined L2 prosodic change for a paralinguistic purpose in L2 production that has been rarely studied to date (exceptions include Li et al., 2013; Maekawa, 2004); if any, there are only a limited number of studies investigating this issue in L2 perception (Chen et al., 2004; Chen, 2005; Fujisaki and Hirose, 1993). Third, we exploited the fact that L2 learners’ L1 and L2 exhibit different uses of the same prosodic property $F_0$. Specifically, $F_0$ is primarily used for a lexical purpose in one language, while in the other, it primarily exhibits a post-lexical function. Our experimental design set a situation in which participants had to coordinate the different uses of $F_0$ in their L1 and L2 to make their L2 prosodic production appropriate. Fourth, we carried out a joint analysis of $F_0$ contours and segmental durations using an advanced statistical analysis method called Functional Principal Component Analysis (FPCA) (Gubian et al., 2011, 2015; Ramsay and Silverman, 2009). FPCA is a mathematical framework that allowed us to compute descriptive and inferential statistical analyses (such as linear-mixed effect regression, henceforth LMER) on whole $F_0$ contours and segmental durations. The most popular method to analyze $F_0$ contours in linguistic research requires phonological labeling using the ToBI system (Beckman et al., 2005; Grice and Baumann, 2002; Venditti, 1997) and measurement of local phonetic variables such as peak heights, pitch ranges, slopes and segmental durations that have to be defined ad hoc as relevant factors to characterize $F_0$ contours. The statistical analysis that follows is then based on those local descriptors. By doing so, there is a risk of a potential loss or distortion of information caused by using inadequate or even biased shape descriptors (e.g., by imposing a fixed number of “relevant” peaks where plateaus or more complex configurations may be found). The application of FPCA on contour analyses reduces this risk. The FPCA tools are in charge of defining and computing data-driven shape and duration descriptors solely based on the statistical properties of the input data, i.e., $F_0$ contours and segmental durations. The output of FPCA is in numerical form combined with a graphical counterpart, the latter showing the meaning of the former in terms of dynamic traits of the input contours. The user (i.e., the linguist who analyzes the data) can continue the analysis by using the numerical output from FPCA as input to further ordinary statistical analysis, in this study LMER.

The rest of the paper is structured as follows: Section 2 provides a literature review and introduces our hypotheses. After that, Section 3 reports methodological facts and steps made for this study. Background knowledge on FPCA is also provided in this methodological section. In Section 4, results of FPCA are translated in a way to show how participants modified $F_0$ and durations. Section 5 provides discussions and summaries of the outcomes, followed by conclusions in Section 6. Mathematical details of FPCA are provided in Appendix A.

2 Hypotheses and related work

From now onwards, the four groups of participants in the present study are named as follows: Japanese L1 speakers and German L1 speakers - speakers who produced
utterances in their L1 - and German L2 speakers and Japanese L2 speakers - speakers who produced utterances in their L2. Note that, in one task, the participants had roles as L1 speakers, and, in a second task, as L2 speakers, i.e., all speakers of both languages performed the same task both in their L1 and in their L2 (see participant details in Section 3.1).

When speakers believe that their utterances were not heard or understood, they reinforce the utterance to try to get the utterances heard or understood. Prosodic adaptations in reinforcement have been widely studied in terms of hyperarticulation (Lombard, 1911; Oviatt et al., 1996, 1998; Pirker and Loderer, 1999; Stent et al., 2008). The term hyperarticulation covers a wide range of adaptations under different environmental factors such as infant-directed speech (Fernald and Simon, 1984; Fernald, 1989; Grieser and Kuhl, 1988; Kitamura et al., 2001), speech to L2 speakers (Sikveland, 2006), speech to automatic speech recognizers (Oviatt et al., 1996; Pirker and Loderer, 1999), or speech in noise, the so-called Lombard speech (Lombard, 1911). On one hand, common phenomena have been clearly manifested across various types of hyperarticulated speech despite the different experimental settings and languages in focus. Previous studies show overlapped results such as longer segmental durations (i.e., slower speech tempo) (Oviatt et al., 1998; Soltau and Waibel, 2000a,b; Stern et al., 1983); exaggeration of pitch contours; variations of $F_0$; a larger pitch range and higher mean $F_0$ (e.g., Fernald, 1989, 1991; Katz et al., 1996; Oviatt et al., 1996; Papoušek, 1992; Trehub et al., 1993); and more rhythmic, musical and stylized intonation (Fernald, 1989; Ladd, 1978; Trainor et al., 1997). Even Chinese-speaking mothers were also found to use higher $F_0$ and wider pitch range in infant-directed speech (Grieser and Kuhl, 1988), showing a tonal language exhibiting the same patterns of acoustic modification as a non-tonal language such as German (see Kitamura et al., 2001, for similar results in Thai). These findings suggest that the exaggerated acoustic characteristics of hyperarticulated speech are universal, although a few studies have reported language-specific modifications as well.

Languages that share similar or common prosodic systems showed similar behavior (see Soltau, 2005, for English and German in hyperarticulation to automatic speech recognizers), while those whose prosody are typologically very different share less common phenomena (see Shochi et al., 2009, for English and Japanese in infant-directed speech). The languages focused on in our experiment, Japanese and German, belong to the latter case, i.e., their respective prosodic systems share few common characters and different language-specific prosodic modifications are expected in both L1 speaker groups and interferences from L1 to L2 are predicted. As with Japanese L1 specific modifications, little room seems to remain to modify prosody for a paralinguistic purpose given the aforementioned lexical restriction in Japanese prosody. Theoretically, the phonological encoding in Levelt’s blueprint of the speaker (Levelt, 1989, 1999) supports the view that the lexical use of prosody outweighs the paralinguistic use of prosody in speech production. The model shows that lexical prosody is assigned to the metrical and segmental spellout before a global prosody is generated in the prosody generator. This means that lexical prosody is processed before the post-lexical and paralinguistic uses of prosody are considered, and that the post-lexical and paralinguistic uses of prosody are generated in a way
that allows no interference to lexical prosody. Translating this assumption into the current experiment, Japanese lexical use of $F_0$ should be maintained even while conveying paralinguistic information.

The question is which aspects of Japanese lexical prosody can still be modified to convey post-lexical or paralinguistic information. The modification of Japanese prosody for a post-lexical purpose has been widely studied in terms of focal prominence and infant-directed speech (e.g., Ishihara, 2011; Kitamura and Burnham, 2003; Nagahara, 1994), while production studies on the modification of prosody to convey paralinguistic information still constitutes the minority of research (Maekawa, 2004; Li et al., 2013; Ofuka et al., 2000; Shochi et al., 2008).

Focal prominence often requires us to examine the relationship among constituents in a whole sentence, whereas our experiment includes an one-word utterance. Still, some of the findings in the previous studies on focal prominence seem to be relevant for our study. For example, Japanese speakers were reported to use a variety of prosodic mechanisms to mark focal prominence, including: local pitch range expansion; dramatic rise at the beginning of the focus constituent to set off the focal constituent; post-focal subordination; and prominence-lending boundary pitch movements, but, importantly, without the phonological modification of a pitch accent (see the review in Venditti et al., 2008). Moreover, analyses on infant-directed speech (Shochi et al., 2009) indicate that Japanese speakers would lengthen an utterance to hyperarticulate speech either by lengthening all morae equally, so that the durational proportions within the word would not change, or by lengthening the word-final morae, an occurrence which has been reported in a variety of languages, which is called utterance-final lengthening (Lehiste, 1972) given that the utterance-final lengthening does not lead to a lexical competition by forming any contrastive pair in terms of length. The utterance-final lengthening has been also found in Japanese, although mainly in spontaneous speech data rather than in laboratory speech (Ota et al., 2003; Port et al., 1987; Warner and Arai, 2001). Warner and Arai (2001) found that the mean duration of the final two morae of the word was generally longer than the mean duration of morae even in speech without hyperarticulation. With respect to the modification of prosody to convey paralinguistic information, Maekawa (2004) and Li et al. (2013) showed that Japanese L1 speakers expressed different emotional statuses by varying maximum and minimum of $F_0$ contours, as well as $F_0$ mean values though notably without changing the phonological form of a pitch accent. Ofuka et al. (2000) investigated how Japanese speakers prosodically convey politeness and showed that $F_0$ of the final vowel of the sentences, i.e., of the question-particle ka, and overall speech rate were significant predictors for the politeness. The sentence-final question particle ka does not carry any pitch accent, thus it does not contribute to a lexical distinction and $F_0$ of the particle can be freely modified. The same is true for the overall speech rate that does not change length proportions of the morae of an utterance, thus it is the cue that can be modified without competing with the rule of Japanese lexical length contrast.

As with German, prosodic modifications for a paralinguistic purpose have been investigated in a series of studies on prosodic features in emotional speech (Burkhardt et al., 2005; Kienast and Sendlmeier, 2000; Paeschke et al., 1999; Paeschke and
Sendlmeier, 2000; Paeschke, 2004), which analyzed varieties of phonetic cues such as $F_0$ range, steepness of $F_0$, duration, spectral tilt and formant in the perception of emotions. Emotional speech generally exhibits higher $F_0$ mean values, larger $F_0$ range, steep $F_0$ pitch accent, longer duration and higher loudness (Paeschke et al., 1999; Paeschke and Sendlmeier, 2000; Paeschke, 2004). These modifications are in line with the aforementioned common observations in hyperarticulated speech across languages. Also in focal prominence, similar modifications in $F_0$ and speech rate have been observed such that if a sentence contains a focus, the reference top line of $F_0$ is raised, provoking a sudden boosting of the pitch accent that correlates with the focused word and longer duration (e.g., Baumann et al., 2007; Braun, 2006; Féry and Ishihara, 2009). Notably, different phonological types of pitch accent are used to mark broad, narrow and corrective foci, i.e., to mark different information structures in German (e.g., Braun, 2006; Féry and Ishihara, 2009), which has not been observed in Japanese due to the lexical restriction of the Japanese pitch accent (Kubozono, 2007). Analyses of German infant-directed speech also showed a greater $F_0$ range and higher mean $F_0$ values and longer durations and more alternating $F_0$ patterns compared to adult-directed speech with different phonological types of pitch accents (Zahner et al., 2015). Apart from the common modifications in hyperarticulated speech across languages (see above), another German-specific prosodic modification in the current study may be the modification of phonological types of pitch accents. It is widely accepted that phonological types of pitch accents in German convey paralinguistic information such as emotions and attitudes of a speaker (e.g., Banse and Scherer, 1996; Baumann et al., 2001, 2007; Grice and Baumann, 2002; Grice and Bauman, 2007; Kohler, 2005, 2008) and specific intonation patterns reflect specific emotions both in speech production and perception studies (e.g., Banse and Scherer, 1996; Baumann et al., 2001; Pierrehumbert and Hirschberg, 1990; Raithel and Hielscher-Fastabend, 2004). For example, a falling contour in German is perceived as more impolite than a rising contour (Baumann et al., 2001). Although previous studies on hyperarticulated speech in German have not reported this modification, we still expect phonological changes in German $F_0$ contours in the repeated utterances produced by German L1 speakers. In particular, participants will be expected to get frustrated in the repetitions because of the communication failure. Therefore, we expect to find more falling pitch contours produced by German L1 speakers in the repetitions.

Based on these previous findings, the following hypotheses for L1 hyperarticulation are stated:

**Hypothesis 1 (Language-specific modifications)** The way in which L1 speakers phonologically modify $F_0$ contours to convey paralinguistic information depends on the use of $F_0$ in their L1.

More specifically, Japanese L1 speakers will not change $F_0$ contours phonologically in the repetitions since a Japanese falling pitch accent is lexically determined. German L1 speakers will vary $F_0$ contours in the repetitions as the $F_0$ contours of the speakers signal one’s attitudinal and emotional states in German. They are more likely to produce a rising contour in the first attempt, whereas in the repetitions this
changes towards a falling contour due to the frustration caused by communication failure.

Secondly, both in Japanese and German L1 data, utterance lengthening and phonetic modifications of $F_0$ (i.e., expansion of local pitch) are expected in the repeated utterances as general phonetic modifications in hyperarticulated speech:

**Hypothesis 2 (Language-independent modifications)** A general utterance lengthening (with a particularly greater utterance-final lengthening) and the expansion of a local pitch range are expected in all speaker groups.

Finally, we will predict the following hypothesis for L2 hyperarticulation:

**Hypothesis 3 (Transfer from L1 to L2)** The difference predicted in Hypothesis 1 holds true in L2 production.

Recent studies on L2 prosody have shown phonetic and phonological interference from an L1 to an L2 in both speech perception and production (e.g., Chen and Mennen, 2008 for L2 English - L1 Italian; Garding, 1981 for L2 French - L1 Swedish and Greek; Jilka et al., 2007 for L2 English - L1 German; Jun and Oh, 2000 for L2 Korean - L1 English; Mennen et al., 2010 for L2 English - L1 Punjabi or Italian; Ueyama and Jun, 1998 for L2 English - L1 Korean or Japanese). Studies on prosodic transfer from L1 German to L2 Japanese and/or from L1 Japanese to L2 German are still scarce and most of these are perception studies (Hayashi et al., 2000; Niikura et al., 2011). Only a few previous studies have focused on prosodic changes for a paralinguistic purpose in L2 production like the current study does (e.g., Maekawa, 2004), and just a few studies investigated this issue in L2 perception (Chen et al., 2004; Chen, 2005; Fujisaki and Hirose, 1993). In absence of related previous experimental work on L2 production, we limit ourselves to generally hypothesize that the phonetics and phonology of L1 prosody will interfere with the acquisition of L2 prosody, so that the use of prosody by L2 speakers differs from the one by L1 speakers. More specifically, German L2 speakers are expected to vary $F_0$ contour in repetitions despite the Japanese lexical restriction. Japanese L2 speakers are predicted to remain faithful to one pattern of $F_0$ contour and will not vary their $F_0$ contours. They will be more likely to produce a falling pitch accent ($H^*+L$) than a rising accent ($L^*+H$) as they are used to doing so in their L1. As general prosodic modifications, both German and Japanese L2 speakers are expected to show local pitch range expansions and utterance-final lengthening in the repetitions.

### 3 Experiment

The experiment reported in this work tested how paralinguistic information affects prosody at the lexical level and vice versa. In a simulated task, participants were instructed to get the attention of a waiter in a crowded and noisy bar by producing the phrase “Excuse me” in Japanese and German repeatedly until they were told that they have succeeded in getting the waiter’s attention. Participants performed the same task both in their L1 and in their L2 language (Japanese and German). The task was designed to elicit the situation in which L2 speakers would have to use a
given prosodic properties ($F_0$ and segmental durations) for a purpose that is different from the purpose that same properties are used for in their L1. We exploited the fact that the phonological form of a Japanese pitch is lexically determined, while the German one is not and can be freely used for a paralinguistic purpose.

3.1 Participants

15 speakers of Tokyo-Japanese who were learners of German (f = 8, m = 7, aged between 19 and 36, mean age = 25.1) and 15 speakers of Standard German who were also learners of Japanese (f = 6, m = 9, aged between 22 and 35, mean age = 28.9) (in total 30 participants) voluntarily took part in the experiment for a small fee. They were all unaware of the purpose of the experiment. None of the participants had any self-reported speech or hearing deficits. Their L2 proficiency was rated using the Japanese Simple Performance-Oriented Test (SPOT-test) (Hatasa and Tohsaku, 1997; Sakai, 2015) that measures Japanese usage ability (max. score = 50, range between 0 – 48, $M = 25.7$, $SD = 16.0$) and the German C-Test (Coleman et al., 1994) that measures general language proficiency (max. score = 60, range between 6 – 60, $M = 41.2$, $SD = 13$). All learners had been learning the L2 at least 2 months (Japanese as L2: range between 2 – 216 months, $M = 50.5$ months, $SD = 58.7$ months; German as L2: range between 2 – 420 months, $M = 111.3$ months, $SD = 115.8$ months). Two of the Japanese L2 speakers have never lived in a German-speaking country (length of stay ranged between 0 – 144 months, $M = 22.4$ months, $SD = 39.0$ months). Seven German L2 speakers had never lived in Japan (length of stay ranged between 0 – 96 months, $M = 27$ months, $SD = 31.7$ months). All participants had learned English as an L2 and none of them had learned any another language with either lexical pitch accents or tones. This series of participant L2 related factors showed that the learners’ L2 learning background/history as well as their L2 proficiency ranged widely from beginners to advanced learners. Based on the analysis of multicollinearity (Belsley et al., 1980), in which two or more predictor variables in a multiple regression model are highly correlated and thus the coefficient estimates of the multiple regression may change erratically and erroneously in response to small changes in the model or the data, only the L2 test scores were used to analyze an effect of L2 proficiency on the performance by L2 speakers in the experiment. Speaking ahead of results, L2 proficiency did not predict their performance, so that the report on the analysis with L2 proficiency is not included in this paper.

3.2 Materials

The target words used in the study were the very frequent Japanese and German words *sumimasen* and *Entschuldigung*. Both words mean *excuse me* and can be used in the same pragmatic context, for instance attention-seeking. *Sumimasen* contains a lexically specified pitch fall associated with the penultimate mora in the word, *se* and four-syllabic (and five-moraic). The word may be optionally realized with an initial low (Gussenhoven, 2004; Vance, 1987), and an example of a typical $F_0$ contour can be seen in Figure 2. The German word *Entschuldigung* is also a four-syllabic
word and the syllable schul is stressed.

![Figure 2: A typical example contour for sumimasen produced by one of the Japanese participants.](image)

### 3.3 Design and Procedure

Following the procedure in Prieto and Roseano (2010), materials participants were presented with descriptions of short scenes and their task was to produce the target word in the given context, which was to attract a waiter’s attention in a crowded noisy bar. The experiment was designed with Microsoft PowerPoint 2008 and was presented on a Macintosh G3 laptop. Each context consists of four slides with the following description written in the participants’ L1 with a picture of a typical situation:

![Figure 3: Example of a slide shown to German participants](image)

**Slide 1** You are in a crowded noisy bar in Japan/Germany. Please call a waiter by saying “Excuse me”.

**Slide 2** He did not hear you. You are a little bit frustrated. Please try it again by saying “Excuse me”.

**Slide 3** He still did not notice it. You are very frustrated. Please try it as the last time by saying “Excuse me”.

11
In this way, we recorded the three attempts (the first attempt and two repetitions) of the same words *sumimasen* or *Entschuldigung*. This 2 (languages) × 3 (attempts) design allowed us to test how the strengthening of one aspect influences the coordination of prosodic properties ceteris paribus.

In addition, 16 filler contexts were provided in which participants were asked to produce Japanese or German short sentences in certain emotional situations such as with anger or irony. The contexts in the participants’ L1 were presented in the first part of the experiment and those in their L2 in the second part. The whole experiment took about 20 minutes. Japanese participants were recruited and tested individually in a quiet room at the Tokyo University of Foreign Studies in Japan and German participants at the Ruhr-University Bochum and Heinrich-Heine-University Düsseldorf, both in Germany. Their responses were digitally recorded onto a computer (44.1kHz, 16Bit) using a unidirectional short-range microphone. After the experiment, they filled out the questionnaire and took the L2 proficiency test.

### 3.4 Data preparation: $F_0$ and time extraction

The prosodic analysis of this work was based on sampled $F_0$ contours and segmental boundaries. We were aware of the fact that vocal expressions of emotion vary along other continuous phonetic dimensions, such as intensity or spectral tilt and voice quality (Banse and Scherer, 1996; Kienast and Sendlmeier, 2000). In the current study, however, we focused on $F_0$ and segmental durations because it was difficult to reliably measure intensity or spectral tilt without the unfixed distance between the speaker’s mouth and microphone and voice quality without perceptual analyses. Moreover, a joint analysis of two functional data would have become difficult for three properties (including intensity or spectral tilt) and voice quality data were not functional data, meaning that these properties were not suitable for a joint functional analysis.

As the first step of the data preparation, the first author annotated the recorded raw data using Praat (Boersma and Weenink, 2011), applying standard segmentation criteria (Turk et al., 2005). Since FPCA requires a continuous input in time, without containing *holes* in the $F_0$ signal, the first fricatives *s* in *sumimasen* and *sch* in *Entschuldigung* were cut off because $F_0$ was absent in these voiceless segments. Moreover, the entire first syllable of *Entschuldigung*, *ent*, was also cut off due to the fact that some participants did not produce it, because, in German, the reduced forms *Tschuldigung*, *Schuldigung* are acceptable forms of *Entschuldigung*. Finally, the missing $F_0$ signal along the second fricatives *s* in *sumimasen* was artificially reconstructed by linear interpolation.

Japanese data were segmented into morae and German into syllables. In detail, internal boundaries were placed as follows: *s | u | m i | m a | s e | n* for the Japanese material, and *ent.sch | ul | d i | gung* for the German material (vertical lines indicate boundaries and strike lines the cut segments). For nasals, the measurement started when the amplitude in the waveform dropped and the waveform showed less high frequency components (drop in high frequency energy in spectrogram). L2 data
were even harder to annotate as L2 speakers showed greater variability in phonetic realizations of segments (for example some Japanese L2 speakers produced [r] in *Entschuldigung* instead of [l]). Such deviant phonetic forms found in L2 speaker data were treated as if they were produced as normative segments. Utterances containing hesitations and wrong words were discarded from the analysis ($N = 8$). In total, 172 utterances were used for the analysis: 86 for *sumimasen* and 86 for *Entschuldigung*.

After the annotation, $F_0$ contours were computed using the $F_0$ tracker available in the Praat toolkit (Boersma and Weenink, 2011). A default range of 70-350 Hz for males and 100-500 Hz for females was used. These ranges were adjusted for specific speakers in order to minimize obvious errors, such as octave jumps.

### 3.5 Functional Principal Component Analysis (FPCA)

In this study, the procedure introduced in Gubian et al. (2015) and in Gubian et al. (2011) was applied. This section provides a brief introduction to FPCA and the main concepts of FPCA in order to understand the workflow of this study, while Appendix A provides mathematical details. Moreover, the code and the data used in this work are available for download at the website maintained by the second author (http://lands.let.ru.nl/FDA).

FPCA is a data-driven technique that produces a parametrization of a set of curves. In this work, curves are $F_0$ contours extracted from a single word, but the same technique can be applied to a longer stretch of speech, like a whole sentence (Gubian et al., 2011; Turco et al., 2011), as well as to other dynamic signals like formants (Gubian et al., 2015). FPCA turns the rich information contained in the curve dynamics into a reduced set of parameters called Principal Component (PC) scores. These scores can then be treated using ordinary statistical modeling methods.

To date, there have been several attempts to apply data-driven (semi)automatic analysis methods to $F_0$ (e.g., the Fujisaki model, Fujisaki, 1992; polynomial equations, Grabe et al., 2007; MOMEL, Hirst and Espesser, 1993 or Tilt, Taylor, 2000). For example, polynomial equations are a way to describe curves including $F_0$ contours in a mathematical expression using variables and constants (Grabe et al., 2007). A fixed “dictionary” of orthogonal shapes (horizontal line, sloped line, parabola and cubic or wave) is used to define the shapes of $F_0$. This method solves the problem simply for short curves, while longer ones may need a more flexible set of basic shapes.

There are two main advantages in using FPCA rather than other methods. First, the parameters that describe the curves are extracted automatically from the data themselves with minimal intervention from the user. Second, the specific procedure used here, joint analysis, allows to capture relevant variation in $F_0$ contours and in segmental duration at the same time.

The input to FPCA is a set of $F_0$ contours, each one paired in the word with the position of the relevant segmental boundaries (e.g., | ul | di | gung | for *Entschuldigung* as shown before). The output of FPCA is a set of scores, with each one associated to a function called *PC curve* (or *PC*), which encodes a systematic variation in
the contour dynamics found in the input data set. For instance, Figure 10 in Section 4 shows the dynamic variation modulated by the PC1 score $s_1$ for the word Entschuldigung. As $s_1$ varies from $-1$ to $1$, $F_0$ contours become flatter and, at the same time, the utterance gets shorter, especially in the last syllable gung. This means that an $F_0$ contour extracted from Entschuldigung with the PC1 score of the value 1 is flatter and shorter than one scoring $s_1 = -1$.

In general, more than one score is needed to capture sufficient information from a data set to characterize the data. PC scores are ordered according to the percentage of explained variance, in the same way as in ordinary PCA. In this work, we extracted the first three PC scores ($s_1$, $s_2$ and $s_3$) both from the sumimasen and from the Entschuldigung data sets (see Appendix A.3 on the distribution of variance across PCs).

Crucially, PC scores are computed disregarding any prior information. In our case, this means that both language groups L1 and L2, as well as utterances from 1st, 2nd and 3rd attempt (cf. Section 3) were pooled together and FPCA did not use this prior information to compute the scores. This allowed the construction of LMER models where language group and number of attempt were used to predict PC scores without introducing any circularity.

4 Results

The data were analyzed with the purpose of characterizing performance differences between L1 and L2 speakers in the attention-seeking task and of examining the results against the hypotheses proposed in Section 2. To this aim, a set of LMER models were built where speaker group (L1 vs. L2) and number of attempt (1st, 2nd and 3rd as an ordered factor) were used to predict numeric quantities related to the shape of $F_0$ contours and to segmental durations of the target word: sumimasen or Entschuldigung. These numeric quantities were the PC scores acquired from the application of FPCA to $F_0$ contours and segmental durations.

The analysis on sumimasen and Entschuldigung was carried out separately and independently. In each case, FPCA was applied on $F_0$ contours and segmental durations jointly (cf. Section 3.5). We considered the first three PCs, in both cases, that collectively explained 68% of the variance for the sumimasen data and 73% of the variance for the Entschuldigung data.

Then, LMER models were built with PC scores (1, 2 or 3) as dependent measures, speaker group and number of attempt as fixed factors, and participant as a random factor including random slopes for the fixed factors (Barr et al., 2013; Cunnings, 2012). P-values were calculated using the Satterthwaite approximation in the R-package lmerTest. Statistical results are reported on the most parsimonious model obtained by eliminating the factors that were not significant, if this did not deteriorate the fit of the model, and by proceeding with backward elimination based on log likelihood ratio tests. Whenever a significant interaction or a main effect was found in the LMER models, plots are shown where the meaning of the predicted PC score is illustrated in terms of $F_0$ contour shape and segmental durational changes. Finally, wherever an interaction between number of attempt and language group was found, multiple pairwise comparisons of each factor were ran using the R-package
lsmeans.

In the following, results from each of the 2 (words) × 3 (PC scores) = 6 LMER models are reported in sequence. To simplify notation, PC scores are always denoted as PC1, PC2 or PC3 scores, or $s_1$, $s_2$ and $s_3$ in plots, irrespective of which of the two independent FPCA analysis they come from.

4.1 Sumimasen

**PC1**: PC1 explained 31.9% of the variance. Figure 4 translates PC1 scores into effects on the shape of time-normalised $F_0$ contours (top) and more durations (bottom). Both plots show the effect of PC1 score as varying between $-1$ and $1$ in 5 regular steps, which matches the variability range of Figure 5.

Figure 4 shows that PC1 modulated a variation of $F_0$ that mainly affected the rightmost part of the curves, corresponding to the more se and n and it had almost no effect on segmental durations. As $s_1$ varied from 1 to $-1$, an $F_0$ contour showed a steeper falling pitch fall in the rightmost part of the curve.

The LMER analysis showed a significant interaction between language group and number of attempt ($\beta = .70, SE = .29, t = 2.4, p < .05$). Figure 5 shows mean PC1 scores and ±1 standard error bars for each attempt in each speaker group and confirms the interaction found in the LMER analysis. The pairwise comparison of number of attempt in each language group showed a significant difference between the first and the third attempt by German L2 speakers ($\beta = .88, SE = .31, t = 2.8, p < .05$). The pairwise comparison of language group in each number of attempt showed a significant difference between Japanese and German speakers in all three attempts ($\beta = 1.81, SE = .24, t = 7.5, p < .001$ in the first attempt, $\beta = .93, SE = .36, t = 2.6, p < .03$ in the second attempt, $\beta = .82, SE = .33, t = 2.5, p < .03$ in the third attempt). While PC1 scores of Japanese L1 speakers were around -1 in all attempts, German L2 speakers' scores decreased from 1 to 0.5 along the increasing number of attempts with a significant difference between the first and the third attempt. The difference between the language groups lied in all three attempts, with a larger difference in the first attempt.

Taken Figure 4 and Figure 5 together, Japanese L1 speakers, whose PC1 scores were around -1 in all attempts, constantly produced $F_0$ contours with a drastic pitch fall (coded as H*-L L-% in the Japanese ToBI, Venditti, 1997), which was a typical phonetic form of a Japanese pitch accent. German L2 speakers, whose PC1 scores decreased from 1 to 0.5 along the increasing number of attempts, produced flat $F_0$ contours in the first attempt (corresponding to the value of 1) and more falling $F_0$ contours in the third attempt (corresponding to the value of 0.5). However, even in the third attempt, the pitch fall produced by German L2 speakers was not as steep as that produced by Japanese L1 speakers and the pitch range produced by German L2 speakers was not as large as that produced by Japanese L1 speakers.

**PC2**: PC2 explained 20.9% of the variance. Figure 6 shows that PC2 modulated both a variation of segmental durations and $F_0$ contours at the same time. As $s_2$ varied from 1 to $-1$, an utterance became longer, especially in the last mora with a nucleus se, and at the same time, its $F_0$ contour exhibited larger pitch range with a steeper pitch fall at the end of the contour. The durational lengthening appeared in
Figure 4: $F_0$ contours (top) and morae durations (bottom) corresponding to PC1 scores $-1, -0.5, 0, 0.5$ and 1 for the word *sumimasen*.

All segments, though the extent of the lengthening was the largest for the last mora with a nucleus *se*. The variation found in $F_0$ contours was a phonetic change that related to this utterance-final lengthening.

The LMER analysis showed a significant interaction between language group and number of attempt ($\beta = -.50, SE = .16, t = -3.1, p < .01$). This interaction is visualized in Figure 7, showing the mean PC2 scores and ±1 standard error bars for each attempt in each speaker group. Further, the pairwise comparison of number of attempt in each language group showed a significant difference between the first and the second attempt and between the first and the third attempt by Japanese L1 speakers ($\beta = .60, SE = .15, t = 3.8, p < .001$ and $\beta = .72, SE = .16, t = 4.5, p < .001$ respectively). The pairwise comparison of language group in each number of attempt showed a significant difference between Japanese and German speakers.
in the first attempt (β = -.56, SE = .24, t = -2.3, p < .03). While PC2 scores of Japanese L1 speakers decreased from 0.4 to -0.3 along the increasing number of attempts, German L2 speakers’ scores were located around 0 and did not change significantly between the attempts.

Taken together, Japanese L1 speakers lengthened their utterances, especially the last mora with a nucleus se, producing a steeper pitch fall with a larger pitch range along the number of attempt. German L2 speakers did not show these phonetic modifications in the repetitions.

**PC3:** PC3 explained 15.6% of the variance. Figure 8 shows that PC3 modulated a variation of $F_0$ contours and segmental durations at the same time. As $s_3$ varied from 1 to −1, an $F_0$ contour became flatter in the initial part of the utterance (corresponding to the more $u$, $mi$ and $ma$) and an utterance became longer accompanied with longer durations of the same initial moras and of the very last mora $n$.

The LMER analysis showed a significant main effect of speaker group (β = 1.13, SE = .19, t = 5.8, p < .001) (see Figure 9). While PC3 scores of Japanese L1 were located above the value of 0 (0.4 on average), German L2 speakers’ scores were located under the value of 0.

By interpreting the visual and statistical analyses together, Japanese L1 speakers produced a typical phonetic form of a contour begin, called initial low (Gussenhoven, 2004; Vance, 1987), while German L2 speakers did not. Moreover, German L2 speakers generally produced longer utterances than Japanese L1 speakers, caused by the longer durations of all morae except for the pitch accented mora se.

### 4.2 Entschuldigung

**PC1:** PC1 explained 27.5% of the variance. Figure 10 shows that PC1 modulated a variation of $F_0$ contours and of segmental durations at the same time. As $s_1$ varied from 1 to −1, an $F_0$ contour showed a steeper pitch fall after the stressed syllable $ul$ toward the final syllable $gung$ and the durations of all syllables became longer with a particularly larger lengthening of the last syllable $gung$. Moreover, the lower the score was, the more likely it was that the utterance contained a falling contour at the end.

The LMER analysis showed a significant interaction (β = .82, SE = .24, t =
Figure 6: $F_0$ contours (top) and morae durations (bottom) corresponding to PC2 scores $-1, -0.5, 0, 0.5$ and $1$ for the word *sumimasen*.

3.4, $p < .01$). Mean PC1 scores and ±1 standard error bars for each attempt in each speaker group are shown in Figure 11, which visualizes the interaction found in the LMER analysis. In order to better understand the nature of the interaction, data were split by *speaker group*. In both speaker groups, there was a main effect of *number of attempt* ($\beta = -1.17$, $SE = .19$, $t = -6.1$, $p < .001$ for German L1 speakers and $\beta = -.37$, $SE = .12$, $t = -3.0$, $p < .01$ for Japanese L2 speakers). The pairwise comparison of *number of attempt* in each *language group* showed a significant difference between the first and the second attempt and between the first and the third attempt by German L1 speakers ($\beta = 1.30$, $SE = .24$, $t = 5.4$, $p < .001$ and $\beta = 1.82$, $SE = .26$, $t = 7.1$, $p < .001$ respectively) and between the first and the third attempt by Japanese L2 speakers ($\beta = .65$, $SE = .23$, $t = 2.8$, $p < .01$). The pairwise comparison of *language group* in each *number of attempt* showed
a significant difference between German and Japanese speakers in the first attempt ($\beta = .64$, $SE = .29$, $t = 2.2$, $p < .03$). Both speaker groups decreased their PC1 scores in the repetitions, but the extent of the decrease was larger for German L1 speakers than for Japanese L2 speakers.

Taken together, the $F_0$ variations shown by German L1 speakers were larger than those shown by Japanese L2 speakers. In the first attempt, German L1 speakers produced contours with a smaller pitch range on the stressed syllable, followed by a stronger rising end with generally shorter segmental durations, while, in the second and the third attempts, they produced a steeper falling pitch accent followed by a flat or falling pitch end with longer utterance durations. Japanese L2 speakers also varied the pitch range and the end of the contour as well as segmental durations in the same tendency that German L1 speakers showed. However, the L2 speakers’ extent of change was significantly smaller than that of L1 speakers. In the repetitions, utterances produced by L1 and L2 speakers became similar owed by the fact that both speaker groups produced $F_0$ contours with a steeper pitch fall at the end.

**PC2:** PC2 explained 23.4% of the variance. The LMER analysis did not show any significant interaction nor main effect.

**PC3:** PC3 scores explained 21.9% of the variance. Figure 12 shows that PC3 modulated a variation of $F_0$ contours and segmental durations at the same time. As $s_3$ varied from 1 to −1, it became more likely that an $F_0$ contour contained a steeper pitch fall followed by a rising end by exhibiting a larger pitch range after the stressed syllable. At the same time, the durations of the stressed syllable $ul$ and the following syllable $di$ became longer.

The LMER analysis showed a significant main effect of speaker group ($\beta = .39$, $SE = .12$, $t = 3.4$, $p < .05$) (see Figure 13). While PC3 scores of German L1 were located under the value of 0, Japanese L2 speakers’ scores were above the value of 0.

Taken together, German L1 speakers produced a steeper pitch fall on the stressed syllable with a larger pitch range which was followed by a rising contour end, while Japanese L2 speakers produced a falling pitch accent followed by a gradual fall toward the end. This phonetic form produced by Japanese L2 speakers made a contour generally flatter than that produced by German L1 speakers. Compared to German L1 speakers, Japanese L2 speakers produced the last syllable $gung$ generally longer,
Figure 8: $F_0$ contours (top) and segmental durations (bottom) corresponding to PC3 scores $-1$, $-0.5$, $0$, $0.5$ and $1$ for the word *sumimasen*.

preceded with shorter syllable durations of *u* and *di*. The durational proportion of all syllables produced by Japanese L2 speakers differed from those produced by German L1 speakers.

5 Discussion

The experiment presented in this work investigated the integration of lexical and paralinguistic prosody in Japanese and German L1 and L2 production. Numerical data output from FPCA were analyzed using LMER to detect the differences between the L1 and L2 groups with respect to $F_0$ contours and segmental durations.

Results showed both L1-dependent and -independent ways to convey paralingu-
guistic information by phonologically and phonetically modifying $F_0$ contours and segmental durations. In the L2 production, clear interferences from speakers’ L1 to L2 were found. The results shown in Section 4 are summarized along the hypotheses proposed in Section 2.

Hypothesis 1 predicted language-dependent prosodic modifications to convey paralinguistic information. Japanese L1 speakers were predicted not to change $F_0$ contours phonologically to convey paralinguistic information due to the lexical restriction of Japanese pitch accent. German L1 speakers were predicted to phonologically vary $F_0$ contours in the repetitions as the variation of $F_0$ contours signals speakers’ attitudinal and emotional states in German. Hypothesis 1 was confirmed by the findings from PC1 scores for *sumimasen* and PC1 scores for *Entschuldigung*. For Japanese L1 speakers, PC1 score for *sumimasen*, which modulated the phonological changes of $F_0$, did not change across repetitions, showing that they constantly produced the same phonological $F_0$ contours. As with German L1 speakers, PC1 score for *Entschuldigung* decreased in the repetitions with a significant difference between the first and the third attempt, which means a higher likelihood to produce a rising contour in the first attempt, i.e., the $F_0$ forms known to signal politeness or friendliness (Baumann et al., 2001), and falling contours in the repetitions that signal frustration. This finding indicates that participants tried to pursue a response in an unsuccessful communication by producing more falling $F_0$ contours. The way German L1 speakers varied $F_0$ contours in the repetitions is in line with the current knowledge about the paralinguistic use of German $F_0$ (e.g., Baumann et al., 2001; Wichmann, 2000).

In agreement with the previous studies (Kitamura and Burnham, 2003; Nagahara, 1994; Li et al., 2013; Maekawa, 2004; Ishihara, 2011; Venditti et al., 2008), our finding shows a strong restriction of lexical pitch accent for a non-lexical use of Japanese prosody and indicates that the lexical prosody dominates the non-lexical one, as supported by the phonological encoding in Levelt’s blueprint of the speaker (Levelt, 1989, 1999).

Hypothesis 2 predicted more exaggerated pitch contours with a larger pitch range and longer utterance durations in the repetitions irrespective of the speaker groups, because these changes are known as general prosodic modifications in hyperarticulated speech (e.g., Fergzsibm, 1964; Fernald, 1991; Katz et al., 1996; Oviatt et al.,...
Figure 10: $F_0$ contours (top) and syllable durations (bottom) corresponding to PC1 scores $-1, -0.5, 0, 0.5$ and $1$ for the word "Entschuldigung."

1998; Papoušek, 1992; Soltan and Waibel, 2000a,b; Stern et al., 1983; Trehub et al., 1993). This hypothesis was also confirmed by PC2 scores for "sumimasen" and PC1 scores for "Entschuldigung." Regarding the Japanese L1 speaker data, their PC2 scores decreased in the repetitions, showing that they produced larger pitch range in the repetitions. This phonetic change occurred together with the utterance lengthening in the repetitions. Japanese L1 speakers lengthened all morae, with a particularly large lengthening on the last mora, with a nucleus $se$, which is known as an utterance-final lengthening that has been reported in a variety of languages (Lehiste, 1972) (though not in a context of repetitions like our study investigated), including in Japanese (Ota et al., 2003; Port et al., 1987; Warner and Arai, 2001). With respect to German L1 speaker data, PC1 scores for "Entschuldigung" modulated both $F_0$ and segmental durations. Their PC1 scores decreased in the repetitions, show-
Figure 11: Mean PC1 scores (as dots) and ±1 standard error bars for the word *Entschuldigung* for each attempt (1st, 2nd and 3rd) in each speaker group (L1 and L2).

ing that they produced rising contours in the first attempt, moving toward falling contours preceded by a larger pitch range on the stressed syllable in the second and third attempts. This change in $F_0$ was accompanied with the change in segmental durations. All segments became longer in the repetitions, with a particularly large lengthening in the final syllable *gung*.

In place of changing an $F_0$ contour phonologically, Japanese L1 speakers raised $F_0$ peaks and produced a steeper pitch fall, which resulted in a local pitch range expansion with a dramatic rise at the beginning of the utterance and a steep fall on the pitch accented mora. This modifications are line with the findings in the previous studies on hyperarticulated speech in Japanese (e.g., Ishihara, 2011; Kubozono, 2007; Nagahara, 1994; Venditti, 2000). As Venditti (2000) argues, Japanese L1 speakers manipulated pitch range to cue many of the same discourse properties which are cued by changing a phonological type of pitch accent in English or German.

Note that, in Japanese, segmental length (proportions of segmental durations) is lexically contrastive, i.e., it exhibits a lexical restriction, as Japanese $F_0$ does. Despite this lexical restriction of segmental length, Japanese L1 speakers modified segmental durations. This was possible because segmental length is determined relatively as proportions within a word, meaning that morae themselves can be lengthened while keeping their relative proportions. As opposed to segmental length, a lexical pitch contrast is perceptually categorical or dichotomous. Therefore, in Japanese, $F_0$ seems to be more restricted than durations for a paralinguistic modification. In our data, the word-final morae were lengthened more than the other morae, because the utterance-final lengthening accelerated further lengthening. The words analyzed in the current study did not form any contrastive pair in terms of length. A further question is whether the durational property could be modified in a similar way when a word builds a minimal pair in terms of length. In such a case, durations would have to be changed in a more careful way than in the case where a word does not build any minimal pair.

As with L2 speakers’ outcomes, Hypothesis 3 predicted that modifications of $F_0$ and segmental durations would be influenced by their L1: German L2 speakers were predicted to vary $F_0$ contours phonologically, even though this change was not linguistically allowed in Japanese. Japanese L2 speakers were predicted not to
Figure 12: $F_0$ contours (left) and syllable durations (right) corresponding to PC3 scores $-1$, $-0.5$, $0$, $0.5$ and $1$ for the word Entschuldigung.

vary $F_0$ contours in the repetitions and they were expected to produce a falling pitch accent rather than a rising accent as the speakers were restricted with the lexical falling pitch accent. Regarding their modifications in segmental durations, utterance lengthening was expected to be found as a general modification as has been reported in the previous studies.

Regarding L2 German speakers, the interaction between speaker group and number of attempt in PC1 scores for sumimasen showed that they varied phonological forms of the Japanese pitch accent despite its lexical restriction. With respect to segmental durations, they did not lengthen the utterances in the repetitions and produced generally shorter utterances than Japanese L1 speakers did. In L2 Japanese speakers’ data, we found the interaction between speaker group and number of attempt in PC1 scores for Entschuldigung. In other words, L2 Japanese speakers
produced a larger pitch range, more falling pitch contours and longer utterances in
the repetitions, though not as much as German L1 speakers did. Since PC1 scores
did not modulate a phonological change of a pitch accent on the stressed vowel $u$, it
can be still claimed that Japanese speakers’ faithfulness to producing a falling
pitch accent was transferred to their L2 production, because Japanese L2 speakers
did not phonologically vary pitch accents, but phonetically. However, as opposed
to their L1 performance, Japanese L2 speakers varied boundary tones in the rep-
etitions. The rising final boundary tone preceded with $H^\ast$ L-, which was found
in our data, is indeed an existing phonological pattern to signal interrogatives in
Japanese (Fujisaki and Hirose, 1993). Asano (2015), who analyzed the same data
manually using the ToBI system, reported that German L1 speakers produced $L^\ast$
-H% (rising pitch accent followed by a rising boundary tone), while Japanese L2
speakers never produced a rising pitch accent, only a rising boundary tone with a
steep pitch rise at the very last part of the utterance. Since the phonological form
found in the final boundary tone was a typical Japanese final boundary tone for an
interrogative sentence, we argue that they may have applied a strategy to produce
Japanese interrogatives for a request (as our task was to ask a waiter to order a
beer) in their German L2 production. To support this argument, it would be in our
further interest to analyze a clear declarative sentence as a control baseline.

In the FPCA analysis, this smaller pitch range and no occurrence of an $L^\ast$
pitch accent in the L2 speaker data were captured as a rather phonetically marginal
variation of rising boundary tones compared to the variation produced by German L1
speakers, thus contributing to the significant interaction in PC1 scores. Interestingly,
these two types of final pitch rising produced by L1 and L2 speakers were classified
by the annotators in Asano (2015) with two distinctive phonological categories, while
FPCA captured these changes as gradual phonetic changes without considering the
perceptual categories. This also indicates that FPCA captures characteristics of $F_0$
contours and segmental duration globally but not locally. The fact that this final
pitch rising produced by L2 speakers did not emerge as a separate PC from FPCA
denotes the main limitation of a data-driven tool like FPCA, where phenomena
are ranked with respect to their quantitative variation without taking perceptual
categories into account.

While the utterances produced by L1 speakers were longer in the repetitions
and the greatest durational change by L1 speakers also took place in the word-final syllable \textit{sen} and \textit{gung}, the comparable durational changes were not observed in the L2 utterances. For \textit{sumimasen}, German L2 speakers did not change segmental durations in the repetitions. For \textit{Entschuldigung}, Japanese L2 speakers did not change them as much as L1 speakers did. We interpret this difference between L1 and L2 data with respect to the utterance lengthening as follows: It might be possible that it was too difficult for L2 speakers to concentrate simultaneously on the appropriate coordination of both $F_0$ and segmental durations in the L2 utterances, given that L2 processing is more demanding due to the increased cognitive load (e.g., Antoniou et al., 2015) and given that hyperarticulation may consume cognitive resources too, while working memory capacity is limited (Cowan, 1995). Consequently, fewer cognitive resources remained available for additional control.

The finding that German L2 speakers did not lengthen their utterance, especially the final morae \textit{sen}, even though they did in their L1 utterances, invites the following interpretations. One possible interpretation is that German L2 speakers stressed the pitch accented morae \textit{sen} by employing the German word stress rule that a final heavy syllable of a word is stressed (Dohmas et al., 2014). When the nucleus vowel with the final syllable \textit{sen} is produced as being metrically stressed (with higher spectral tilt), it may be articulatory difficult to lengthen this syllable. The difficulty in achieving Japanese mora-timing rhythm in an L2 when the L1 of a learner is a stress-timed language has been reported in previous studies (e.g., Kondo, 2005; Masuko and Kiritani, 1992). For example, Kondo (2005) reported that it was difficult for English L1 speakers to separate the acoustic features of $F_0$ and vowel duration with the manifestation of lexical stress, and thus they often fail to manipulate and control two acoustic features of stress in English, i.e., vowel duration and $F_0$ increase independently in L2 Japanese production. Only a few very advanced English speakers showed a good mora timing control in Japanese, greatly increasing the $F_0$ for accented morae without increasing mora durations. Another interpretation is that duration, in German, is a less dominant prosodic property than $F_0$ to convey a paralinguistic meaning. This may be contrary to Japanese, where duration is interpreted as the more dominant property than $F_0$, because the perception of a segmental duration within a word is relative to the proportions of other segmental durations of the word and thus may be less restricted than $F_0$. Japanese L2 speakers lengthened their utterances in their German L2 utterance because the durational property was a more dominant prosodic property in their L1 (Japanese) to convey paralinguistic information. Since German L1 speakers also varied durations (as a less dominant property to change in German), Japanese L2 speakers’ attempts to vary durations still appeared to be successful. In the case of German L2 speakers, they changed a more dominant prosodic property in their L1 (German) and varied $F_0$ in the same way as Japanese L2 speakers did in their L2 utterances. Differently to the case for \textit{Entschuldigung}, however, Japanese L1 speakers did not change $F_0$, so the German L2 speakers’ attempts turned out to be not successful. Our proposed argument is that L2 speakers are more prone to change a more dominant prosodic property in their L1 to convey a paralinguistic meaning, i.e., duration for Japanese and $F_0$ for German.

Similarly to durational changes that were expected to be found in all speaker
groups, it was predicted that a larger pitch range in the repetitions would be found in all groups. This phonetic change was only observed in both L1 speaker groups, whereas L2 speakers either did not modify it or their change was smaller than L1 speakers’ one. This finding that L2 speakers did not produce higher $F_0$ peaks and larger pitch ranges in the repetitions is in line with Urbani (2012) and Wennerstrom (1994, 1998). The papers reported that the utterances produced by L2 speakers bore a narrower pitch range than those produced by L1 speakers. One possible interpretation for this finding is that L2 speakers generally feel unsure when speaking in an L2 and that they thus do not boldly expand pitch ranges. Another interpretation relates to the aforementioned argument that L2 speakers had difficulties in appropriately coordinating $F_0$ and segmental durations simultaneously. Since German L2 speakers produced the Japanese pitch accent with a metrical stress accompanied with a generally shorter segmental duration, they might have not had enough temporal space to expand the pitch range. This appears to be especially true for German L2 data, because durations and $F_0$ are dependent on each other in German (Beckman, 1986; Fujisaki et al., 1986).

The results in this study were discussed reflecting that the participants tried to convey paralinguistic information, i.e., increased frustration due to the communication failure, although such attitudinal changes of the participants are difficult to measure quantitatively (Nolan, 2008). It may be argued that the participants might have not felt stressed even in the second or third attempt, so that there was actually no “increased frustration” to convey in the repetitions. First of all, we believe that the participants did not necessarily have to actually feel stressed to accomplish the task appropriately, but it was sufficient for them to follow the task instructions and produce the words as if they had felt stressed. In the repetitions, participants were explicitly instructed to imagine being more frustrated due to the unsuccessful communication. Findings from the data support the claim that they did so: The multiple pairwise comparison of number of attempt showed that a difference between the attempts in each language group mainly lied in the contrast between the first and the second or the third attempt, but not between the second and the third attempt. More specifically, the lengthening of the utterances were generally observed in the second and the third attempts, which is known as a general prosodic adaptation in hyperarticulated speech. This can be evidence for signaling the increased frustration in the repetitions. Further, rising pitch accents and boundary tones were mostly found in the first attempt both in the German participants’ L1 and L2 utterances. Such rising contours are known to signal politeness and friendliness (Baumann et al., 2001). Our data therefore suggest that German participants signaled politeness and friendliness in the first attempt by producing rising contours. In the repetitions, participants tried to pursue a response with falling $F_0$ contours which may sound less polite (Baumann et al., 2001). Finally, Japanese L1 speakers showed a significantly higher pitch followed by a steeper pitch fall exhibiting a larger pitch range in the second and the third attempt compared to the first attempt. All these findings were phonetic and phonological changes due to the hyperarticulation caused by the pragmatic manipulation. One could still argue that even the first attempt already included a certain speech act (e.g., calling) (e.g., Searle, 1969) and participants hyperarticulated the utterance in the first attempt, because the task
description at the beginning of the task indicated that the bar was noisy. Taken the finding that a main difference between the attempts lied in the contrast between the first attempt and the repetitions, the degrees of the reinforcement between the first attempt and the repetitions were still significantly different, and the nature of the hyperarticulation in the first attempt and in the repetitions was different. Nevertheless, in order to clearly rule out this argument, it would be worth including a control baseline condition in future work in which participants produce an utterance without eliciting any intention to hyperarticulate an utterance.

Furthermore, the multiple pairwise comparison of speaker group further showed that a difference between the two language groups mainly lied in the first attempt. This result lends itself to the interpretation that the L2 speakers had difficulties in producing appropriate L2 prosody independently from conveying the attitudinal change (= increased frustration). With respect to L1 and L2 speakers' PC scores in the repetitions that often did not significantly differ from each other, we still do not feel confident in drawing the conclusion that Japanese L1 speakers and German L2 speakers or German L1 speakers and Japanese L2 speakers applied the same strategies to hyperarticulate utterances in the repetitions. The reason for this is that both groups showed different patterns of changes across the attempts. Therefore, their PC scores in the repetitions can happen to be so similar, still being related to different strategic reasons.

What may be further included as future cases in the line of research would be the potential relation of acoustic strategy and oral communicative skills in L2. The L2 proficiency tests used in our study were claimed to measure general language proficiency including oral skills (Oller, 1976). The reason why we did not find any effect of L2 proficiency on the learners' performance in this study may be due to the large range in the learners' L2 proficiency level and their further variable learning backgrounds. In our future research, more controlled learner groups with clearly different L2 proficiency levels may be tested.

Finally, our findings are especially noteworthy as they were found in highly frequent words, in both Japanese and German, (“excuse me”) that L2 speakers should have encountered very often before. Despite a rich amount of input in their L2, they still failed to produce appropriate forms of L2 prosody, suggesting the difficulties to acquire L2 prosody. These findings motivated us to conduct further perception and production experiments to determine the source of these deviant L2 production forms in L2 perception, storage and production (Asano, 2014, 2016).

One of the novelties of this study was the use of the advanced statistical method FPCA to analyze $F_0$ data. This semi-automatic data analysis method was expected to reduce the number of subjective decisions involved in assigning a ToBI category or in a local phonetic measurement. Moreover, a further advantage in the use of FPCA was in enabling us to analyze whole $F_0$ data by capturing the most distinctive important characters to define contour differences. Finally, the joint analysis of $F_0$ and segmental durations made it possible to observe interactive changes in $F_0$ and segmental durations, which is otherwise difficult to investigate in a single manual analysis.

The outcomes obtained from FPCA were in line with our own previous investigations with manual annotation using a ToBI and manual measurement of durational
changes (Asano, 2016, Chapter 2). In addition, FPCA provided more insightful findings, showing relationships between the change in $F_0$ and segmental durations providing numerical data to conduct statistical analyses on whole contours. FPCA could identify the most distinctive and important phonetic and phonological characters for a typical Japanese utterance (e.g., an initial low and a falling pitch accent) (Gussenhoven, 2004; Vance, 1987) without any prior information about their importance. A further advantage of FPCA was found in analyzing L2 data. L2 data have been otherwise reported to be difficult to annotate using the ToBI (Asano, 2016; Asano et al., 2016) due to its more creative and dynamic characteristics of an interlanguage containing deviant and unexpected phonetic and phonological realizations (Selinker, 1972). Indeed, several difficulties have been reported when using ToBI annotation for the analysis of L2 data (Asano, 2016; Asano et al., 2016). First, one selected ToBI system (either of an L1 or of an L2) can hardly cover all deviant variabilities of L2 data that represent the interlanguage characteristics. A language-specific ToBI system does not account for the deviant L2 realizations with respect to L1, so unexpected contours might be hard or even impossible to describe using a single ToBI system of either the speakers’ L1 or L2. Second, the annotators’ language backgrounds may influence the annotation. Should an annotator be an L1 listener of German when annotating German utterances produced Japanese L2 speakers? Or is it better when the annotator is a Japanese L1 listener? These difficulties and problems reported in the previous annotation analyses demonstrate the appeal of employing a different analysis method, as we did in the current study by using FPCA.

However, FPCA has its limitations, like any method does. As briefly discussed before, FPCA does not take perceptual categorical thresholds into account. It is therefore important to compare the statistical results and the visual information provided in the FPCA plots with original data to correctly interpret results and to ensure that a statistical analysis corresponds to an interpretation based on our linguistic knowledge. Moreover, FPCA does not differentiate perceptually important local parts of continuous $F_0$ data. In future works, it may be worth exploring modifications of FPCA where user-defined weights can be applied to single parts of a contour to tell the system which parts are perceptually more decisive for comparing contours. In this way, small but perceptually crucial differences in a portion of $F_0$ contour may not emerge from automatic modeling.

6 Conclusions

The coordination of lexical and paralinguistic prosody in L1 and L2 hyperarticulation was examined using the joint analysis of FPCA on $F_0$ and segmental durations. Lexical prosody outweighed paralinguistic prosody and L1 speakers conveyed paralinguistic information without breaking the restriction of lexical prosody. Moreover, different dominance orders of prosodic properties may exist in Japanese and German according to which prosodic property was more likely to be modified in hyperarticulated speech. L2 speakers were more prone to change a more dominant prosodic property in their L1 to convey a paralinguistic meaning, i.e., segmental durations for Japanese and $F_0$ for German. The interferences from L1 to L2 showed that it
was difficult for L2 speakers to coordinate $F_0$ and segmental durations in hyper-articulated speech. The results reported in this study are especially telling as the analyzed words were very frequent words, confirming the difficulties encountered in the acquisition of L2 prosody despite a rich amount of L2 input.
References


A FPCA for prosodic data

This appendix provides an overview of the data processing method used in this article for the analysis of $F_0$ contours. The method is based on FPCA, which is one of the data analysis algorithms available within the Functional Data Analysis (FDA) framework (Ramsay and Silverman, 2009; Ramsay et al., 2009). Since the method is described in Gubian et al. (2015) and Gubian et al. (2011) in detail, this section focuses on a concise exposition of the main concepts and an outline of the mathematical backbone. Readers interested in applying FDA to their own data are invited to visit the website maintained by the second author (http://lands.let.ru.nl/FDA), from which instructions and codes can be downloaded. The rest of the appendix is divided in three sections that correspond to the main steps of the procedure, namely smoothing, landmark registration and FPCA, respectively.

A.1 Smoothing

Smoothing transforms a contour composed of discrete samples into a continuous curve represented by a mathematical function (this is what “functional” refers to in FPCA or FDA). The target function is chosen from a set of possible functions specified by the user. In the case of $F_0$ contours, which can assume a wide range of shapes, it is customary to adopt B-splines as the general function set (De Boor, 2001). A B-spline is a sequence of polynomial curves that, summed together, approximate the desired contour. By varying the number of polynomial components of B-splines and a parameter influencing the curve smoothness, the user controls how close to the original discrete samples the continuous curve will be, or, in other words, the user controls the temporal resolution of the continuous representation (Ramsay and Silverman, 2009). An example is shown in Figure 14, in which the smoothing curve was chosen in such a way to eliminate some microprosodic detail present in the sampled contour. In general it is important to find a good compromise between “smoothing too much”, which would delete relevant information from the original contours, and “not smoothing enough”, which would leave irrelevant detail that makes the extraction of global trends harder in the following steps.

A.2 Landmark registration

The analysis of prosody is generally based on descriptions of the $F_0$ movement in relation to the underlying segments, e.g., syllables or morae. However, FPCA produces a model based on physical time, where the position of the relevant segmental boundaries is different from one contour to another. If we used the original contours as input to FPCA, at the end of the analysis we would not be able to distinguish contour shape variations that are a consequence of a variation of segmental durations (e.g., a peak occurs always inside the same segment, but the preceding segments vary their duration) from variations of pitch alignment (e.g., a peak occurs before or after a given segmental boundary).

In order to interpret the results of FPCA in terms of the segmental structure underlying the contour, a nonlinear time warping called landmark registration is applied to the original (smoothed) $F_0$ contours that synchronizes all corresponding
Figure 14: Example of a smoothed $F_0$ contour. Dots represent $F_0$ samples obtained from the pitch tracker available in Praat. The curve is a B-spline. This contour is extracted from a realization of the word *Entschuldigung*, where the first syllable was discarded (cf. Section 3.4). The y-axis reports $F_0$ values in semitones after the global mean value was subtracted (thus corresponding to the zero level).

Landmark registration alters the time axis of a curve in such a way that landmarks, i.e., segmental boundaries, move from their original position to a new position specified by the user; usually the average position computed on the whole curve set is chosen as common location of landmarks. The operation is carried out automatically and it is only based on the time position of the selected landmarks where the latter can obtained either from the manual annotation, e.g., carried out in Praat, or automatically using a speech recognition software (e.g., see Turco and Gubian, 2012). The time warping involved in landmark registration is guaranteed to preserve the qualitative aspects of curves, i.e., all rise and fall movements are preserved in the original order and level, no jumps or ruptures are produced. Figure 15 shows some examples.

The output of landmark registration is a new set of curves where all corresponding segments have the same duration, which means that the information on the differences in segmental durations is discarded. Gubian et al. (2011) proposed a way to represent the original durations by means of functions (curves) and to incorporate this representation in FPCA. Let $h(t)$ be the time warping curve produced by landmark registration that specifies the mapping between normalized time $t$ and original time $h(t)$ for a specific normalized curve $F_0(t)$ (bottom left panel in Figure 15). This curve contains all the information necessary to reconstruct the original $F_0(t)$ contour (top left panel in Figure 15) from its registered version (top right panel in Figure 15). An equivalent representation of $h(t)$ that is more convenient for the statistical analysis tools that follow is obtained by

$$r(t) = -\log \frac{dh(t)}{dt}.$$ (1)
Figure 15: Three $F_0$ contours randomly chosen from the 86 realizations of the word *Entschuldigung*, where the first syllable was discarded (cf. Section 3.4). Top left: $F_0$ contours before landmark registration. Top right: $F_0$ contours after landmark registration. Bottom left: time warping functions $h(t)$ describing landmark registration for the three contours. Bottom right: time warping expressed as relative log rate $r(t)$ as in Equation (1). Three different line styles are used across the four panels in order to identify each contour. Dots on the curves mark the position of the landmarks (here syllabic boundaries $|ul| di |gung|$), and vertical dashed lines show the position of the registered landmarks, which correspond to the average position computed on the entire set of 86 realizations of this word.

The curve $r(t)$ represents the log of the relative rate of realization of the utterance with respect to its normalized version. According to Equation (1), if a segment (i.e., anything between two consecutive landmarks) is realized two times faster than in the normalized version, then $r(t)$ will read values around $\log_2 = 0.7$ along the time interval corresponding to that segment; vice versa, values around $\log_2^{-1} = -0.7$ will be reached for segments pronounced at half rate (i.e., double duration). The bottom right panel in Figure 15 shows curves $r(t)$ obtained from $h(t)$ in the bottom left panel. The advantage of using $r(t)$ instead of $h(t)$ is that the former represents proportional variations in duration linearly. Fortunately, since Equation (1) is com-
pletely reversible, once we obtain results in terms of $r(t)$ we can convert them to $h(t)$ values, which correspond to landmark positions, or directly to differences between $h(t)$ values, which correspond to segmental durations. This will be shown in the next section.

A.3 FPCA

The actual statistical analysis on $F_0$ contours is carried out by FPCA. The input to FPCA consists of a set of (multidimensional) curves, which, in our case, are $(F_0(t), r(t))$ pairs, where $F_0(t)$ is represented in normalized time and $r(t)$ was defined in Equation (1). The output is a compact model composed of a few curves, called Principal Components (PCs), which combined in appropriate amounts reproduce every input curve, with some approximation. Formally, every curve $(F_0(t), r(t))$ is decomposed as:

$$F_0(t) = m_{F_0}(t) + s_1 \cdot PC1_{F_0}(t) + s_2 \cdot PC2_{F_0}(t) + \cdots $$

$$r(t) = m_r(t) + s_1 \cdot PC1_r(t) + s_2 \cdot PC2_r(t) + \cdots $$

Functions $m_{F_0}(t)$ and $m_r(t)$ are the mean curves, i.e., the curves obtained by computing the mean value of all input curves at each instant in (normalized) time in the $F_0$ and in the $r$ dimension, respectively. Functions $PC1_{F_0}(t)$ and $PC1_r(t)$ are the first PC curves in the $F_0$ and in the $r$ dimension, respectively, and the same holds for the second and further PCs. Finally, $s_1$, $s_2$, etc. are the so-called PC scores, which are coefficients that differ for every input curve and that determine the amount of correction to the mean $(m_{F_0}(t), m_r(t))$ that each PC should contribute in order to approximate a given curve $(F_0(t), r(t))$. PCs are ordered by decreasing percentage of explained variance. Crucially, Equation (2a) and Equation (2b) share the same PC scores. This means that PCs act jointly on the mean $(m_{F_0}(t), m_r(t))$, thus jointly capturing variations in the shape of time-normalized contours and in their segmental durations (Gubian et al., 2011). All PCs are independent from one another, i.e., the shape and duration variations described by PC1 are not correlated with those described by PC2, PC3, etc. This means that given an input curve and its $s_1$ score, we cannot predict its $s_2$ and further scores.

In order to mitigate the possibility that the variance of one of the two dimensions dominates the other, a weight is applied to the curves based on the ratio between the variance of the dimensions (here omitted to simplify notation). However, the dynamics of $F_0$ curves remain rather different from that of log rate curves. This has the consequence of producing an unexpected imbalance of explained variance, e.g., a modest variation in segmental length (coded as log rate) may dominate a local yet (linguistically) more relevant variation in $F_0$ contour shape. Thus, it is advisable to look at the first few PCs, since interesting variations are not necessarily those that explain more variance.

PC scores can be used as variables in ordinary statistical analyses. In this article, they were used as independent variables in linear mixed-effect models. Once a prediction is obtained in terms of PC scores, this has to be translated back into its meaning in terms of $F_0$ contours. Suppose we build a model based on $s_1$ and we obtain a predicted value $\hat{s}_1$ from it. By substituting $s_1 = \hat{s}_1$ in Equation (2)
we obtain two predicted curves $\hat{F}_0(t)$ and $\hat{r}(t)$. While the former is of immediate interpretation, the latter is not. If we are interested in the durations of segments delimited by the landmarks, we can compute them by inverting Equation (1). Let $d_i$ be the normalized duration of the $i$-th segment, which spans the interval $[t_{i-1}, t_i]$ on the normalized time axis $t$ (i.e., $d_i = t_i - t_{i-1}$). The duration $\hat{d}_i$ in the original time corresponding to the warping induced by $\hat{r}(t)$ is given by:

$$\hat{d}_i = \int_{t_{i-1}}^{t_i} \exp(-\hat{r}(t)) dt.$$  \hspace{1cm} (3)

Results presented in this paper are based on Equation (3), which allows to represent the effect of PC scores in terms of plain segmental durations (cf. Figures 4, 6, 8, 10 and 12 in the main text).