Rapid and Smooth Pitch Contour Manipulation

Michele Gubian[1], *Yuki Asano*[2], *Salomi Asaridou*[3,4], *Francesco Cangemi*[5]

[1]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2]Department of Linguistics, University of Konstanz, Germany
[3]International Max Planck Research School, Nijmegen, The Netherlands
[4]Donders Institute for Brain, Cognition and Behavior, Nijmegen, The Netherlands
[5]University of Toulouse II, France and University of Cologne, Germany

m.gubian@let.ru.nl, yuki.asano@uni-konstanz.de,
s.asaridou@donders.ru.nl, fcangemi@uni-koeln.de

Abstract

Speech perception experiments are often based on stimuli that have been artificially manipulated, e.g. to create hybrids between two given prosodic categories. Tools like the widely used PSOLA re-synthesizer available in Praat provide the user full editing control on the shape of pitch and intensity contours, as well as on local relative speech rate of recorded utterances. However, high level experimental specifications, e.g. "generate a number of pitch contours whose shapes are gradual transitions between two reference contours", are not easily translated into a sequence of low level manipulation operations. Often, the viable solution is the manual stylisation of contours, which drastically reduces the degrees of freedom, but at the same time can introduce unwarranted alterations in the stimuli. In this paper we introduce a method, implemented in a software tool interfaced with PSOLA, that automatically translates high level specifications into low level operations in a principled way. No manual editing in the form of contour stylization or otherwise is required. The tool enables the rapid generation of manipulated stimuli of desired properties, while it guarantees that acoustic feature alterations are always smooth. Three use cases demonstrate the efficacy of the tool in real experimental conditions.

Index Terms: stimuli manipulation, perception experiments, prosody

1. Introduction

The artificial manipulation of natural speech is common practice in the preparation of stimuli for perception experiments. A number of speech processing software tools, like the PSOLA (Pitch Synchronous Overlap Add Method) re-synthesis tool available in Praat [1], offer the possibility to modify the shape of f_0 and intensity contours extracted from recorded utterances, and to selectively alter segment duration. Speech scientists use these tools to investigate the perceptual effect of those features in isolation, e.g. by keeping the original f_0 contour and varying segment duration or vice versa. These tools have opened new possibilities for the experimental verification of phonological theories of prosody and intonation. Manipulated resynthesized stimuli are also being used in combination with brain imaging in the investigation of the mental processes involved in language acquisition and learning. However, changing prosodic parameters in isolation can give rise to stimuli that sound unnatural if constraints on the co-variation of the parameters are violated. Therefore, there is a need for embedding manipulation tools in a work bench that makes it possible to vary parameters in combination and to quickly generate large numbers of stimuli that can then be assessed auditorily to remove unnaturally sounding tokens.

In this work we focus on two typical scenarios involving f_0 contour manipulation. The first one is the creation of f_0 contours whose shapes gradually vary between two reference contours. By analysing subjects' responses (e.g. through reaction times) to stimuli whose f_0 contours present hybrid characteristics between two known categories, it is possible to make inferences on the perceptual space, which in general is not linearly mapped on the acoustic space. The second scenario involves the creation of stimuli where one or more speech features are imported, or 'transplanted', from other stimuli. An example is the creation of hybrids that do not exist in a natural language, like the creation of stimuli that combine spoken words in a non-tonal language with f_0 contours extracted from words spoken in a tonal language (cf. Section 4.2).

Although stimuli manipulation procedures are widely used, a closer analysis of the common practice reveals a number of operational limitations. In the case of gradual modification of f_0 contours between two reference shapes, the creation of intermediate shapes using default Praat tools requires two operations, namely (i) aligning corresponding segmental boundaries in time, and (ii) stylizing the reference f_0 contours using straight line segments. The manipulation is carried out by changing the position of the junction points (usually only one) in the stylized contours using a graphical editor or a script (e.g. both options available in Praat). For example, in [2] artificial mixtures of two Mandarin tones are generated from a three-points stylization of each tone and by gradually moving the point in the middle. Similarly, in [3] the shape of a pich rise is changed from concave to convex by imposing a three-points stylization and by varying the middle point height, and in [4] a modulation between two more complex shapes (dip and hat) is obtained by moving more than one point at the same time. While segmental alignment is performed in order to preserve the anchoring of f_0 movements to the segmental material, f_0 contour stylization is not necessarily justified by theoretical or experimental reasons. On one hand, stylization may help in reducing the complexity of a prosodic model by isolating simple shape features. On the other hand, there is always the risk of losing important detail, e.g. the type of curvature (concave or convex) of a rising gesture [5]. Stylization is often carried out manually, which entails empirical judgement and time consuming procedures. Semi-automatic stylization procedures exist, which are claimed to preserve all perceptual properties of the original f_0 contour (see [6] for an overview). However, those faithfullness guarantees refer to the so-called 'close copy' of a contour, thus they do not necessarily extend to manipulation.

While full f_0 contour grafting does not necessarily involve stylization, which is sometimes applied nonetheless (e.g. in [7]), segmental alignment remains a requirement. Ideally, the time warping involved in the alignment should alter the utterance structure as little as possible, in order to minimise the risk of introducing unwanted perception effects. Praat provides manual editing facilities for the selective manipulation of segment duration, and scripted procedures are also available, like [8] or the one used in [9]. To our knowledge, the available scripts apply duration changes locally on each segment, which may introduce noticeable discontinuity effects whenever alteration values of adjacent segment durations differ too much.

In this paper we present a contour manipulation method that eliminates the limitations described above. The proposed method (i) implements smooth deformation of f_0 and intensity contours in order to align them to given segmental boundaries, and (ii) provides a transparent way to combine two or more contours in desired proportions, e.g. for the creation of hybrids or for averaging. The procedures are automatic and controlled by the user through a few paramenters, like the degree of 'elasticity' of contour deformation. This method improves the experimental procedures involved in stimuli manipulation in two ways. First it minimises the risk of introducing unwanted alterations in the original speech samples. Second, it provides for the automatic generation of stimuli in large quantities, enabling the rapid examination of several experimental conditions at design time.

The rest of the paper is structured as follows. In Section 2 the method is described in its main principles, which are adapted from techniques applied in Functional Data Analysis [10]. In Section 3 we give a brief description of the software tools that carry out all the necessary operations. In Section 4 we report three use cases where the method has been applied in real experimental conditions. Finally, in Section 5 we draw conclusions.

2. Method

2.1. Smoothing

We illustrate the basic principles of the proposed method by referring to the manipulation of f_0 contours; similar considerations apply for intensity contours. The first operation to be carried out on the input data is called *smoothing*, which transforms a sampled f_0 contour into a continuous curve represented by a mathematical function of time $f(t)$. This function is constructed by combining a set of so called basis functions such that the combination fits the sampled data. In the case of features like f_0, whose contours in time can assume arbitrary shapes and do not present periodicity, it is customary to adopt B-splines as basis [11]. An example of smoothing is shown in Figure 1. The user can control the degree of smoothing through a number of parameters (see Section 3).

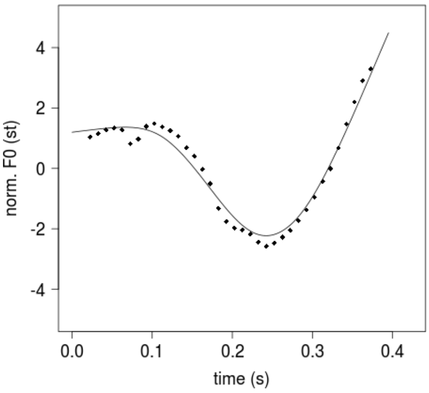

Figure 1: *Example of a smoothed f_0 contour. Dots represent f_0 samples obtained from the pitch tracker available within Praat. The curve is a B-spline. This contour is extracted from a realisation of a three-syllabic word. The y-axis reports f_0 values in semitones after the global mean value was subtracted (corresponding to the zero level).*

Once contours are represented by functions of time, expressing combinations of them becomes trivial. For example, to create hybrids between two base contours A and B, the arithmetic operation $(1-\alpha)\cdot f_A(t)+\alpha\cdot f_B(t)$ produces the desired combinations in proportions controlled by the parameter α.

2.2. Landmark registration

Suppose A and B are realisations of the same three-syllabic word. The operation $(1-\alpha)\cdot f_A(t)+\alpha\cdot f_B(t)$ defined above combines values of contours A and B at corresponding points in time. However, the inevitable segment duration differences between the two realisations would mix shape traits belonging to different syllables, which would blur the timing relation between f_0 movement and segmental content.

This problem is solved by a convenient transformation applied to the time axis that alters each contour in such a way that corresponding segmental boundaries get aligned in time. This operation, called *landmark registration*, is carried out automatically and it is based on the position of each boundary (landmark) on each of the input contours. The time warping carried out by landmark registration guarantees that the qualitative aspects of the curves are not altered. Moreover, the local speech rate alterations are spread gradually throughout the entire contour, which minimises discontinuity effects. Figure 2 shows the effect of landmark registration on an f_0 contour extracted from a three-syllabic word, where the syllable boundaries are shifted to a desired position.

2.3. Manipulation and re-synthesis

Here we illustrate how to combine smoothing and landmark registration in order to create of a set of stimuli whose f_0 contours are combinations of two base contours A and B, extracted from two realisations of the same word or phrase in two different conditions (e.g. yes-no question vs. statement); analogous steps are required in other manipulation schemes.

First, f_0 contours are extracted from utterance A and B,

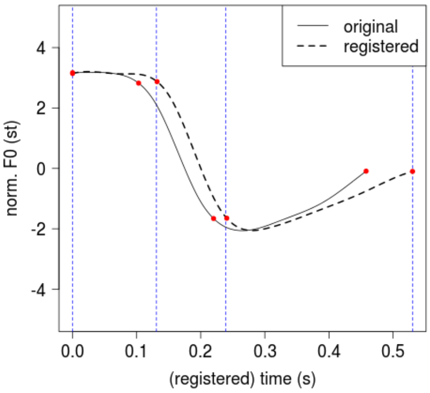

Figure 2: Example of landmark registration of a smoothed f_0 contour extracted from a realisation of a three-syllabic word. Dots show the position of syllable boundaries, vertical dashed lines the position where the boundaries are going to be shifted by landmark registration. The solid curve is the original f_0 contour, the dashed curve the contour after registration.

and all relevant landmarks, like syllable boundaries, are marked (e.g. on a Praat textgrid). Then f_0 contours are smoothed and turned into functions, $f_A(t)$ and $f_B(t)$, which have different duration and whose landmarks are not synchronised yet. Suppose we want to carry out re-synthesis on the recording of utterance A, let us call it the *base* utterance. Before mixing f_0 contours of A and B we have to synchronise utterance B on the landmarks of the base. Thus, landmark registration is carried out on the boundaries of utterance B by imposing a time warp that aligns its boundaries on those of the base A. This is internally represented by a warping function $h_{B\rightarrow A}(t)$. After this, function $h_{B\rightarrow A}(t)$ is applied on $f_B(t)$ to obtain a different function $f_B(t_A)$, which has (qualitatively) the shape of $f_B(t)$ but is aligned with the landmarks of the base A. At this point we can create a number of mixtures of the form $f_\alpha(t) = (1-\alpha)\cdot f_A(t) + \alpha \cdot f_B(t_A)$, say for $\alpha = 0, 0.2, 0.4, \ldots, 1.0$, where the value $\alpha = 0$ will produce a stimulus that should be identical to the original A and will be employed in the experiment in order to control for the re-synthesis effect, as well as being a useful sanity check for the re-synthesis. Finally, all the $f_\alpha(t)$ are converted into PitchTiers and used in Praat PSOLA re-synthesizer to modify the shape of the f_0 contour of the base utterance A.

3. Software

The software to carry out all the operations described above consists of a main R script [12] and a number of auxiliary R and Python scripts (www.python.org). The package is developed and maintained by the first author and is available for download from his website [13] (direct link [14]). The core functionalities are based on the `fda` library [15], with minor modifications. The software accepts Praat formats as input (e.g. TextGrids) and produces output also in Praat formats (e.g. PitchTiers) or wave files by calling Praat. The main script is not intended to be executed in a single call, because the procedure is composed by a cascade of operations, some of which require the user to set a number of parameters, like those controlling smoothing, whose outcome has to be checked by plotting (cf. Figure 1). A simple expedient has been devised in order to alleviate the problem of having gaps in f_0 contours due to voiceless sounds. This is a hindrance when such a contour is transplanted on speech material where voiced sounds occupy the f_0 gap, as reported in [16]. The input contour is smoothly interpolated by (automatically) padding extra samples at a level computed by averaging neighbour sample values.

4. Use Cases

In this section we describe three perception experiments, conducted by the second, third and fourth author, respectively, where stimuli manipulation was carried out using the method and the software introduced above.

4.1. Perception of pitch and length cues in Japanese as a second language

The second author is conducting a study on the influence of first language (L1) in the discrimination of pitch and length in a second language (L2). A number of AX (same-different) discrimination tasks were designed based on German as L1 and Japanese as L2, where German participants were either learners of Japanese or not. In Japanese, which has both consonantal and vocalic length contrasts, pitch appears to be a secondary cue for length [17, 18]. For German participants, whose L1 does not have the consonantal length contrast, the discrimination of length is expected to be harder for consonants than for vowels [19], while the role of pitch in discrimination of length has not been studied yet.

Six non-sense disyllabic words were created that respect Japanese phonotactics and differ from each other in manner of articulation and voicing of the medial consonant. Each word is created in three versions, which differ either in the duration of the first vowel or in the duration of the second consonant, resulting in a singleton (CVCV), a geminate (CVCCV) and a long-vowel (CVVCV) as counterparts (e.g., /punu/, /pu:nu/, /pun:u/) [20].

All of the tokens were produced either with a lexical pitch accent on the first syllable (High-Low) or with no pitch accent at all (High-High). Each token was recorded six times by the same L1 speaker of Japanese, in order to present different tokens to the participants. Five segmental boundaries were considered: /C/V/C/V/, /C/V/CC/V/ or /C/VV/C/V/.

AX tasks were designed to measure discrimination in one cue, either pitch or duration, while keeping the other constant. In order to keep segment duration constant across stimuli, landmark registration was applied as illustrated in Section 2.2, where the 12 realisations of a given segmental pattern (e.g. six /pu:nu/ realisations for each of the two pitch patterns) were aligned on the average time location of those boundaries across realisations. Conversely, to keep pitch constant across stimuli in the duration contrast condition, a cascade of two landmark registrations was applied as follows. An average f_0 contour was created by first aligning the $N = 18$ contours $f_i(t)$, $i = 1, \ldots, N$ (i.e. six realisations times three duration patterns) on the landmarks of one of them, say the first one, obtaining $f_i(t_1)$. This allowed to compute a time-aligned average contour $f_A(t_1) = \frac{1}{N}\sum_i f_i(t_1)$. This average shape was re-aligned on the segmental timing of each of the N realisations, obtainining N pitch-normalised contours $f_A(t_i)$, which were eventually imposed on the recorded stimuli. The naturalness of the stimuli was highly satisfactory in both manipulations. Moreover, f_0 padding was successfully used in order to accommodate for gaps due to voiceless segments (cf. Section 3). The results of the AX tasks are being collected and analysed at the time of writing of this paper.

4.2. Dutchinese

A perception experiment is being conducted by the third author with the purpose of investigating neural and genetic correlates of sound learning performance. The study was performed using Dutch native speakers as subjects. We planned to use a phonetic contrast that would be unknown and difficult enough for Dutch natives while being ecologically valid. We chose the pitch contrasts used in the four Mandarin tones. Following the paradigm used by [21], we opted for Dutch-Chinese hybrid stimuli, i.e. words that respect Dutch phonotactic rules but with Mandarin tone contours superimposed on them. By using hybrid stimuli we could create minimal quadruplets where we manipulated f_0 whilst keeping all the other variables (e.g. word duration, intensity, vowel length, production rate etc.) constant. Twenty-four pseudowords with a consonant-vowel-consonant (CVC) structure were created (e.g. /ket/, /ba:f/, /nal/, /be:m/). We recorded eight Dutch native speakers reading aloud the list of those 24 CVC pseudowords. Similarly, we recorded eight native speakers of Chinese uttering the word /mi/ on four Mandarin tones.

Manipulation was applied for the grafting of f_0 contours from the word /mi/ to the pseudowords. The boundary between /m/ and /i/ was disregarded, since the duration of /m/ was always so short compared to /i/ that no difference could be noticed by considering the f_0 contour shape either as starting from the onset of /m/ or of /i/. On the other hand, we had to investigate how to align the Chinese tone contours on the CVC tokens. In /C/V/C/ there are four boundaries, we called them 1, 2, 3, 4. Using the software described in Section 3, the Mandarin tone contour from the word /mi/ was applied on each pseudoword in 4 different ways: from boundary 1 to 3, 1 to 4, 2 to 3, and 2 to 4. Landmark registration was applied, which in this case was equivalent to a linear time compression or expansion, because only two landmarks were present. Depending on the nature of the consonants (voiced or voiceless) and tones, some combinations were expected to be better than others. The full application of this scheme produced thousands of stimuli, one for each pseudoword, tone, Dutch speaker, Chinese speaker and alignment criterion. From those, a subsample of 384 tokens was extracted and used in a rating study, whose purpose was to find the combinations that would result in the highest tone identification as well as highest naturalness ranking when judged by native Mandarin speakers.

The naturalness of the sound was very satisfactory with the exception of three out of eight speakers whose voices sounded less natural in some of the tokens. The identifiability of tones by native speakers partially suffered from the experimental requirement of excluding all cues except for pitch, as different tones tend to have different durations in Mandarin. This problem mostly affected tone 3, which is the longest in duration. The general trend of the rating study revealed that overlap from boundaries 1-3 and 1-4 were preferred compared to 2-3 and 2-4 although that varied as a function of the Dutch-Mandarin speakers pitch agreement, the specific phonemes, and the tone in question.

4.3. Tempo and sentence modality in Italian

In languages such as Italian and its regional varieties, sentence modality contrasts, e.g. the opposition between declaratives and yes-no questions, are conveyed through prosodic means alone. The intonational aspects, expressed phonetically by f_0 contours and their synchronization with segments and syllables, have been thoroughly studied in production and perception (e.g. [22] for Neapolitan Italian). Recent studies, however, point to the existence of consistently produced differences in segmental durations as well (e.g. [23, 24], and [25] for Neapolitan Italian), but no evaluation of their perceptual role had been provided yet. In order to study the perceptual role of these temporal aspects, it is crucial to manipulate both f_0 contours and durational patterns. If tempo is relevant in the perception of sentence modality contrasts, we expect listeners to react differently to stimuli featuring the same f_0 contour but different durational patterns. Moreover, we can also expect that the effect of temporal manipulations will be stronger if the intonational cues are made unavailable or ambiguous.

The fourth author tested these hypotheses by having 26 subjects participate in a forced-choice identification task based on 18 manipulated stimuli [26]. These were created by manipulating two base stimuli, namely the sentence *Danilo vola da Roma* ('Danilo takes the Rome flight') read as a Question (bQ) and as a Statement (bS) (notation coherent with [26]). For both stimuli, we extracted phone durations (dQ, dS) and f_0 contours (fQ, fS). Then we defined an acoustically Ambiguous durational pattern (dA) as the average of corresponding phone durations in dQ and dS, and an acoustically ambiguous f_0 contour (fA), as the average of fQ and fS. This conceptual scheme was operationalised by applying landmark registration and averaging as explained in Section 2. For example, to obtain an Ambiguous f_0 contour in the duration pattern of a Statement, first apply landmark registration on fQ it its base timing ($fQdQ$) to synchronise it on the phone pattern dS, obtaining contour $fQdS$. Then compute the average $(1-\alpha)\cdot fQdS+\alpha\cdot fSdS$ with $\alpha = 0.5$ to obtain $fAdS$, i.e. an ambiguous f_0 contour in the timing of the Statement base utterance. Finally $fAdS$ can be directly applied to bS, or also to bQ, provided that it is first manipulated by applying the duration transform from dQ to dS. This scheme allowed for the generation of 18 stimuli, i.e. the combination of three levels (Q, S and A) on two factors (f_0 and duration) applied on both base utterances.

The transformation of a base stimulus into its opposite sentence modality was extremely successful in that subjects' responses to original questions ($bQfQdQ$) were not significantly different from responses to statements re-synthesized as questions ($bSfQdQ$), and the same happend for the $bSfSdS$ - $bQfSdS$ pair. On the other hand, the creation of stimuli with ambiguous intonation did not yield the expected results, since subjects exhibited a strong question-bias. This is because in the absence of established knowledge on the warping of perceptual space for utterance-long stimuli, function averaging was accomplished by combining f_0 contour with equal weights $\alpha = 0.5$. Apparently, acoustical ambiguity does not always result in perceptual ambiguity. Despite this limitation, we were able to conclude that listeners do not seem to use temporal information when categorizing stimuli as questions or statements.

5. Conclusions and future work

In this paper we have presented a method for the rapid and effective manipulation of f_0 and segmental duration values aimed at the re-synthesis of stimuli for speech perception experiments. The method provides an automation layer between the level of specification of segmental alignment constraints and contour linear combinations on one hand, and the lower level provided by state-of-the-art editors, like the one available in Praat (PSOLA). The effectiveness of the method was illustrated by three use cases, where it was successfully applied in real experimental conditions. The software that implements the method is available for download and use [13]. Several extensions of the software are possible, for example an explicit mechanism for expressing rigid contour shift, which is a way to create timing variants (e.g. [27]).

6. References

[1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.3.42) [computer program]," *online: http://www.praat.org/*, 2013.

[2] R.-X. Yang, "The phonation factor in the categorical perception of mandarin tones," in *in Procceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011.

[3] E. Dombrowski and O. Niebuhr, "Shaping phrase-final rising intonation in german," in *in Proceedings of the 5th International Conference on Speech Prosody, Chicago, Illinois, USA*, 2010.

[4] G. I. Ambrazaitis and O. Niebuhr, "Dip and hat pattern: a phonological contrast of german?" in *in Proceedings of the 4th International Conference of Speech Prosody, Campinas, Brazil*, 2008.

[5] E. Dombrowski and O. Niebuhr, "Acoustic patterns and communicative functions of phrase-final f0 rises in german: Activating and restricting contours," *Phonetica*, no. 62, pp. 176–195, 2005.

[6] D. Hermes, "Stylization of pitch contours," in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, I. Augurzky P.and Mleinek, N. Richter, and J. Schliesser, Eds. Berlin, New York: De Gruyter (= Language, Context, and Cognition 3), 2006, pp. 29–62.

[7] O. Niebuhr, "The signalling of german rising-falling intonation categories - the interplay of synchronization, shape, and height," *Phonetica*, no. 64, pp. 174–193, 2007.

[8] H. Quené. (2011) Software tools - adjustdurpitch.praat. [Online]. Available: http://www.let.uu.nl/ Hugo.Quene/personal/tools

[9] P. B. de Mareüil and V. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, pp. 247–267, 2006.

[10] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis - 2nd Ed.* Springer, 2005.

[11] C. de Boor, *A Practical Guide to Splines, Revised Edition*. Springer, New York, 2001.

[12] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[13] M. Gubian. (2013) Functional data analysis for speech research. [Online]. Available: http://lands.let.ru.nl/FDA

[14] ——. (2013) Rapid and smooth pitch contour manipulation - software tool. [Online]. Available: http://lands.let.ru.nl/FDA/papers/rapid_smooth_manipulation.zip

[15] J. O. Ramsay, G. Hookers, and S. Graves, *Functional Data Analysis with R and MATLAB*. Springer, 2009.

[16] S. Winters and M. G. OBrien, "Perceived accentedness and intelligibility: The relative contributions of f0 and duration," *Speech Communication*, vol. 55, 2013.

[17] K. Kinoshita, D. M. Behne, and T. Arai, "Duration and f0 as perceptual cues to japanese vowel quantity," in *Proc. of the International Conf. on Spoken Language Processing, Denver*, 2002, pp. 757–760.

[18] H. Kubozono, H. Takeyasu, M. Giriko, and M. Hirayama, "Pitch cues to the perception of consonant length in japanese," in *Proc. of the17th International Congress of Phonetic Sciences Hong Kong*, 2011, pp. 1150–1153.

[19] H. Altmann, I. Berger, and B. Braun, "Asymmetries in the perception of non-native consonantal and vocalic length contrasts," *Second Language Research*, vol. 28, no. 4, pp. 387–413, 2012.

[20] D. M. Hardison and M. Motohashi-Saigo, "Development of perception of second language japanese geminates: Role of duration, sonority, and segmentation strategy," *Applied Psycholinguistics*, vol. 31, no. 01, pp. 81–99, 2010.

[21] B. Chandrasekaran, P. D. Sampath, and P. C. M. Wong, "Individual variability in cue-weighting and lexical tone learning," *Journal of Acoustical Society of America*, vol. 128, no. 1, pp. 456–465, 2010.

[22] M. D'Imperio and D. House, "Perception of questions and statements in Neapolitan Italian," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds., Rhodes, 1997, pp. 251–254.

[23] J. Ryalls, G. Le Dorze, N. Lever, L. Ouellet, and C. Larfeuil, "The effects of age and sex on speech intonation and duration for matched statements and questions in French," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2274–2276, 1994.

[24] V. van Heuven and E. van Zanten, "Speech rate as a secondary prosodic characteristic of polarity questions in three languages," *Speech Communication*, vol. 47, no. 1, pp. 87–99, 2005.

[25] F. Cangemi and M. D'Imperio, "Local speech rate differences between questions and statements in italian," in *Proceedings of the 17th International Congress of Phonetic Sciences*, W. Lee and E. Zee, Eds. Hong Kong: City University of Hong Kong, 2011, pp. 392–395.

[26] F. Cangemi and M. DImperio, "Tempo and the perception of sentence modality," *Laboratory Phonology*, no. 4(1), in press, 2013.

[27] O. Niebuhr and H. R. Pfitzinger, "On pitch-accent identification – the role of syllable duration and intensity," in *in Procceedings of the 5th International Conference on Speech Prosody, Chicago, Illinois, USA*, 2010.