DIAPIX-FL: A symmetric corpus of problem-solving dialogues in first and second languages

Mirjam Wester[1], *M. Luisa García Lecumberri*[2], *Martin Cooke*[3,2]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK
[2]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[3]Basque Foundation for Science, Bilbao, Spain

m.wester@inf.ed.ac.uk, garcia.lecumberri@ehu.es, m.cooke@ikerbasque.org

Abstract

This paper describes a corpus of conversations recorded using an extension of the DiapixUK task: the Diapix Foreign Language corpus (DIAPIX-FL) . English and Spanish native talkers were recorded speaking both English and Spanish. The bidirectionality of the corpus makes it possible to separate language (English or Spanish) from speaking in a first language (L1) or second language (L2). An acoustic analysis was carried out to analyse changes in F0, voicing, intensity, spectral tilt and formants that might result from speaking in an L2. The effect of L1 and nativeness on turn types was also studied. Factors that were investigated were pausing, elongations, and incomplete words. Speakers displayed certain patterns that suggest an on-going process of L2 phonological acquisition, such as the overall percentage of voicing in their speech. Results also show an increase in hesitation phenomena (pauses, elongations, incomplete turns), a decrease in produced speech and speech rate, a reduction of F0 range, raising of minimum F0 when speaking in the non-native language which are consistent with more tentative speech and may be used as indicators of non-nativeness.
Index Terms: L1-L2, DIAPIX

1. Introduction

Talkers modify their speech depending on the environment and their interlocutor, for example, by using Lombard speech in noisy environments [1], more simplified speech when addressing infants [2, 3], and a slower speech rate when talking to non-native listeners [4, 5] (see [6] for a recent review). Research is ongoing to create speech technology systems that are also capable of modifying their output to be more intelligible to listeners depending on the context, e.g., in the presence of noise [7–10] or for specific subgroups of listeners, such as elderly or non-native listeners [11, 12]. In order to modify the speech appropriately, information about the context is required, for instance, the presence of noise, the type of noise or what type of listener is using the system. In this paper, our focus is on the latter, and specifically we explore the problem of distinguishing native and non-native speakers of the target language.

A body of research has focussed on finding which suprasegmentals determine accentedness. Similarly, our work explicitly avoids segmental cues since these are more likely to reflect the pronunciation features of particular first language (L1) - second language (L2) pairings rather than universal characteristics of foreign accent. Moreover, a study by [13] found that around half the variance in oral proficiency and comprehensibility has its basis in suprasegmental features. The main factors that have been found to influence perceived non-nativeness are: speech rate [5, 14, 15], pausing [16], stress [5], fundamental frequency (F0) [13], F0 range and voice quality [17]. However, this is not an exhaustive list, and different studies find more or less influence of these various factors, possibly because many of the factors found to relate to non-nativeness can be highly variable across individuals.

Most of the above cited studies consider only the non-native speech of their subjects: their L1 speech characteristics are generally not taken into account. An exception is the study by Riazantseva [16], which is one of a handful of papers [18, 19] that consider talkers' L1 speech production in addition to their non-native speech production. To our knowledge, there are no studies that contrast two languages by having non-native and L1 productions in both languages. For the current study, we recorded English and Spanish native talkers speaking both English and Spanish, i.e., the native and non-native talker is the same person. This enables us to attempt to separate the factor of language (English or Spanish) from that of speaking in an L1 or L2. Section 2 describes this corpus: DIAPIX-FL (Diapix Foreign Language).

We are interested in discovering if there are suprasegmental features which encode non-nativeness and if so, which are the most prominent ones. We also want to find out to what extent global suprasegmentals and conversational phenomena in the L2 are talker-specific and/or language specific and which are L2 traits. To answer these questions, an acoustic analysis of native (N) and non-native (NN) speech from DIAPIX-FL was carried out (Section 3) as well as an analysis of the use of turn types and their properties in the two languages (Section 4).

2. The DIAPIX-FL Corpus [1]

2.1. Materials

Picture materials from the DiapixUK task were used to elicit spontaneous speech [20]. In this task, two people are recorded while solving a spot-the-difference task. Each participant is given a version of the same cartoon picture that is different in small ways. They have to cooperate to find the 12 differences without being able to see each other's picture. Picture-based elicitation is an ideal method for capturing L1 and L2 speech from the same talkers as it ensures the use of an identical task in each language.

The full set of DiapixUK materials consists of three themes (Beach, Farm, Street), with four pairs per theme. We selected two of the four pairs per theme to use in our recordings. The

[1]The DIAPIX-FL corpus is freely available at http://datashare.is.ed.ac.uk/handle/10283/346

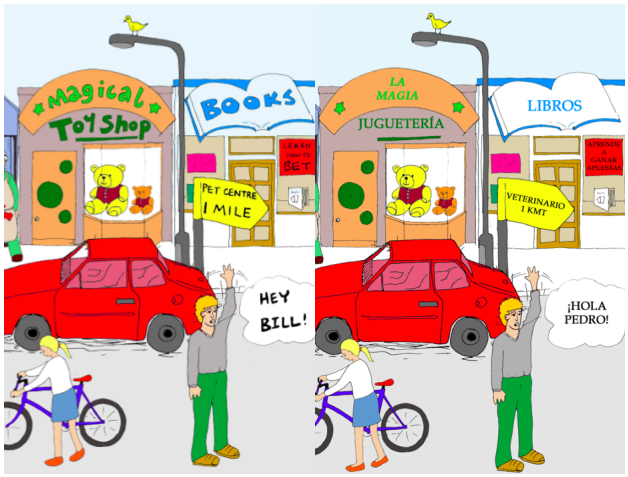

Figure 1: *Fragment of the DIAPIX street scene. Left: original English version; right: DIAPIX-FL Spanish version.*

pictures are cartoons but involve some (English) text, for example on street signs, posters or shop fronts. All the text in the pictures was translated from English into Spanish, creating a set of six English pictures and six Spanish pictures. A fragment of one of the scenes is shown in Figure 1.

2.2. Participants

Speaker pairs were recruited at the University of Edinburgh and at the University of the Basque Country in Vitoria. Per site, six pairs were recorded, i.e., twelve speakers: ten female and two male. The Edinburgh participants were all native English speakers in their 2nd year of university studying Spanish. Similarly, the Vitoria participants were all native Spanish/Basque speakers studying English in their 2nd year at university. To ensure the level of proficiency across the speakers and languages was comparable we selected students with a CEFR level of B2/ C1 for their foreign language [21].

2.3. Recording setup and procedure

Each pair of participants was seated in the same room at a desk with a divider between them so they could not see each other's picture. Their speech was recorded via close-talking microphones and a table microphone. Each pair of speakers was asked to talk through six pictures in total, three in each of English and Spanish. Half of the pairs started with the English pictures and ended with the Spanish pictures. Pairs saw a different version of the same scene in each language. Picture order was balanced across speaker pairs. Before the actual recording started the participants were given the DiapixUK training picture pair of a park scene to familiarize themselves with the task. Participants were remunerated for their time and effort.

2.4. Transcription

The recordings were orthographically transcribed by native speakers of each language and cross-checked by bilingual speakers. Table 1 shows the extra symbols used to annotate aspects such as pausing and incomplete words, as well as suprasegmental characteristics such as elongations and extralinguistic features such as inbreaths. We refer to the segments of speech, non-speech and other events as turn types and distinguish between filled pauses (e.g., "uh", "um", "er"), unfilled pauses which take place during a talker's turn, and silence on the part of the listener when the interlocutor is talking.

Table 1: *Transcription symbols and descriptions.*

?	transcriber unsure of utterance
#	non-speech
@	external noise
*	incomplete word
:	elongation
%	filled pause
<	inbreath
+	unfilled pause
-	silence (because of interlocutor speaking)
$	code-switching

3. Acoustic analysis

3.1. Subset selection

A subset of the DIAPIX-FL Corpus was chosen to support an analysis of potential changes in parameters related to F0, intensity, spectral tilt and formants which might result from speaking in an L2. This analysis was based on material produced by the 10 English and 10 Spanish female talkers. To avoid very short utterances (e.g., back-channels), only those contributions longer than 1.4 s were selected. Since the total amount of speech produced by each talker showed considerable variation, a fixed overall duration of 60 s was used for each talker. These minimum and total duration thresholds were chosen based on the amount of speech produced by the least voluble talker. Segments were chosen at random to avoid bias in the selection of material from any specific phase of the recording. The subset of DIAPIX-FL used in the acoustic analysis thus consisted of 40 1-minute composites of speech (one sample for each of the 20 talkers speaking in their L1 and their L2).

3.2. Speech parameter estimation

The following parameters were extracted from the composite speech signals: (i) intensity level; (ii) mean, minimum, maximum and standard deviation of F0; (iii) spectral tilt, based on linear regression of 1/3-octave band energies; (iv) formant frequencies F1-F3; (v) energy in the frequency bands 1-2, 1-3 and 1-4 kHz; (vi) proportion of voiced frames. Formant differences F2-F1 and F3-F2 and ratios of narrowband to wideband energy were also computed. Intensity, F0, and formant frequencies were computed using Praat [22] while the remaining parameters made use of custom Matlab code.

All parameters apart from spectral tilt, intensity level and band energies were extracted in 10 ms frames and subsequently reduced to a single value for each 60 s speech sample for statistical analysis. Specifically, to remove pitch-halving and other F0 errors, a two-component mixture of Gaussians was fitted to the F0 distribution estimated across each individual 60 s of speech using the expectation-maximisation algorithm, and the mid-point between the two component F0s used as a lower cut-off, below which F0 estimates were deemed unreliable. In cases where the estimated mean F0s were closer than 50 Hz, a single Gaussian was fitted. From the reliable F0 values, robust estimates of mean, standard deviation, minimum and maximum F0 were extracted after removing outliers (defined as values more

than 1.5 times the inter-quartile range below or above the first and third quartile boundaries). A similar procedure was used to produce robust estimates of mean F1, F2 and F3 frequency.

3.3. Results

Separate two-factor ANOVAs with within-subjects factors of nativeness of language being spoken (speaking in their L1 or L2) and between-subjects factors of L1 (English or Spanish) were computed for each acoustic parameter. Figure 2 depicts all outcomes which showed statistically-significant effects.

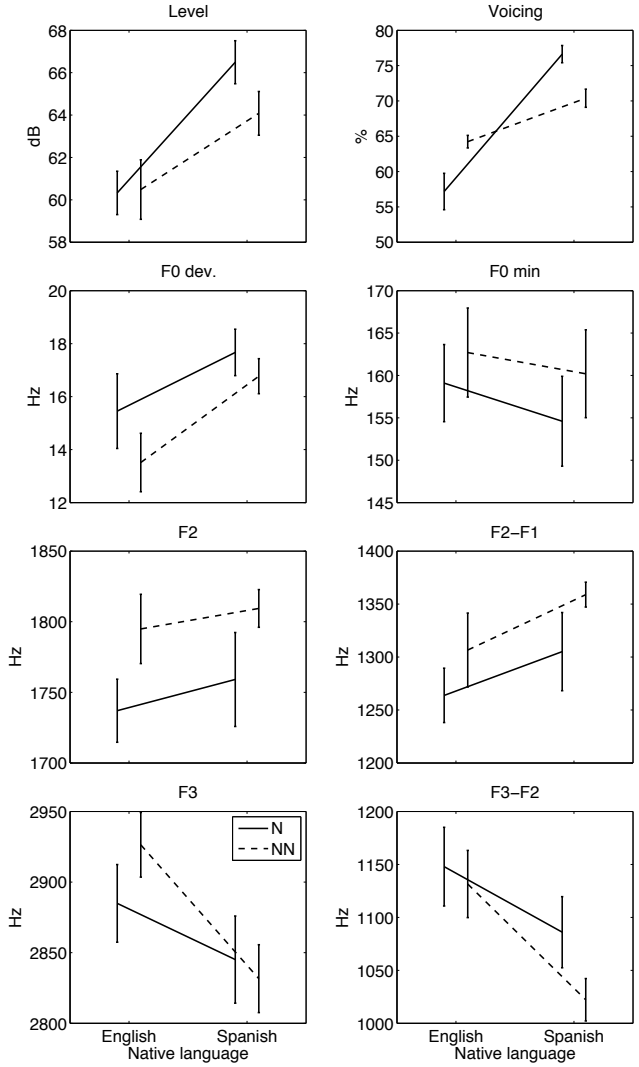

Figure 2: *Acoustic parameters showing significant effects. Here and elsewhere error bars indicate ± 1 standard error, 'N' stands for native and 'NN' for non-native.*

Level Spanish speakers spoke with around 2 dB lower intensity when talking in their non-native language [$p < .01$]. Due to microphone and recording level differences at the two sites, the difference between L1s cannot be interpreted.

Voicing The percentage of voiced frames differed [$p < .001$] across the two languages, with Spanish having around 40% more than English when spoken as an L1. When spoken non-natively, the voicing ratio moved some way in the direction of the language spoken natively.

F0 dev. The variation in F0 decreased when speaking non-natively [$p < .01$], with a tendency [$p = .067$] for larger variations for Spanish speakers.

F0 min. The lower limit on F0 showed a slight increase when talking non-natively [$p < .05$].

F2 freq F2 increased by around 50 Hz ($\approx 3\%$) for both L1s when spoken non-natively [$p < .01$].

F2-F1 Commensurate with the change in F2, the F2−F1 distance increased when speaking non-natively.

F3 freq For English speaking Spanish, F3 increased by about 40 Hz ($\approx 1.4\%$). The interaction approaches significance [$p = .06$].

F3-F2 For Spanish speaking English, F3−F2 decreased [$p < .05$].

No significant changes in F0 mean, F0 maximum, F1, spectral tilt, nor any of the three narrow-to-wideband energy ratios were observed.

4. Turn types and their properties

4.1. Turn types

Using the categories shown in Table 1, the effect of L1 and nativeness of speech production on turn-types was analysed based on ANOVAs with the same structure as used for the acoustic analysis of the previous section. For the analyses of this section the entire recording for each of the 20 female talkers was used. Figure 3 (left) summarises those turn-types exhibiting statistically-significant effects of one or more factors or their interaction. In each case, the value plotted is the percentage of that type of turn in the corpus.

The three types of turn depicted in Figure 3 (left) make up nearly 98% of the corpus, and all show both L1 and nativeness effects. The L1-based differences may be due to differences in productive competence between the two cohorts rather than L1-specific factors. Apart from the number of turns containing speech, L1 and nativeness did not interact. Speaking in a non-native language led to a larger proportion of turns that were silent pauses [$p < .001$] or filled pauses [$p < .001$]. Unsurprisingly, the proportion of turns containing speech was correspondingly reduced [$p < .001$].

4.2. Speech rate, elongations and incomplete turns

The right column of Figure 3 displays the number of words, elongations and incomplete words per minute of speech.

The rate of native and non-native speech cannot be used directly as the structure of each language is likely to influence the 'canonical' speech rate. This is true whether the speech rate is measured in phonemes, syllables or words per unit time. To get around this, we computed a language adjustment factor by comparing the speech rate for the two languages spoken by native talkers. Since the speakers in the two languages are doing the same task (and they are similar in age, education etc.), this appears to be a justifiable method for the measurement of intrinsic language-dependent speech rate differences. A simple words per minute (WPM) measure was used. When spoken natively, the English cohort produced around 235 WPM, somewhat more than the Spanish cohort's 217. Consequently, *all* English (i.e.,

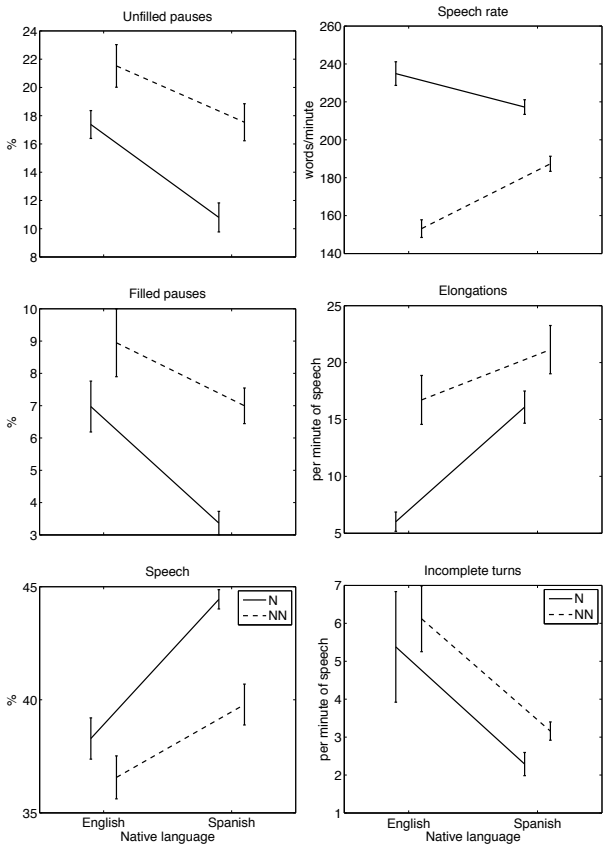

Figure 3: *Left column: turn types showing significant effects. In each case the percentage of turns of the given type is plotted. Right column: speech rate, elongations and incomplete turns.*

spoken by both natives and non-natives) was reduced by a factor of 0.93 in order that the mean 'normalised' speech rate is the same for natives speaking their own language. Based on this normalised measure, a non-native proportional change in WPM can be calculated, i.e., expressing their non-native speech WPM as a fraction of their native speech WPM. Talkers exhibited a reduction in all cases, ranging from 61% to 95%. The mean WPM for English talkers speaking Spanish was 153, a reduction to 70.7% of the normalised native speech rate. For Spanish talkers speaking English this figure is 173, a reduction to 80.0%. Both of these decreases are significant [$p < 0.001$].

Similarly, both groups produced far more elongated words per minute of speech when speaking non-natively, although, as for the speech rate measure, the increase was larger for the English cohort. The number of incomplete words per minute of speech was marginally affected by the nativeness factor [$p = .063$] and also showed larger cohort differences.

5. Discussion

One of the primary motivations for this study was to assess the degree to which talking in a non-native language is similar – at least in terms of its effect on acoustic properties – to other modified speech styles such as clear [23], Lombard [1] or foreigner-directed speech [4, 5]. Our acoustic analysis suggests that L2 speech shares relatively few properties with these modified styles. While the observed increase in F2 is common to these speech styles, other significant changes go in the opposite direction (e.g., reduced intensity level and decrease in F0 variation) and for other parameters which typically exhibit an increase (e.g., F0 mean, mid-frequency energy) no changes were seen.

However, the acoustic and turn type analyses reveal commonalities with findings on polite speech [24] and hesitancy [25]. Grawunder and Winter [24] report that formal speech exhibits a decrease in level and F0 range, an increase in the number of fillers, and a decrease in speech rate compared to more informal speech. Similarly, F0 range reduction and F0 raising tend to convey hesitancy [25,26]. Both politeness and tentativeness involve less categorical styles of speech and in this respect it is not surprising that non-native speech shares features with both.

One interesting outcome was the relationship between the proportion of voiced frames and L1/L2. Spanish possesses in the main a CV structure so that consonants appear less frequently per syllable than in English. Additionally, English rhythm often results in considerably shortening of vowels in unstressed syllables. Other differences between the two languages, such as plosive VOT, contribute to the overall larger proportion of voicing in Spanish. These proportions are clear in our L1 data with native Spanish showing around 30% more voicing than English. When speaking as an L2, talkers modify their voicing in the direction of the target language, but only with a partial approximation to L2 phonological norms. Given these talkers' level of proficiency, this is likely to be due to a certain residual amount of L1 sound transfer but also to some sub-phonemic properties such as VOT or vowel length still being influenced by the L1 system.

The turn type analysis identified four main indicators of non-nativeness, largely replicating earlier findings mentioned in the Introduction using our new paradigm. First, the time spent producing pauses – both filled and unfilled – is longer in non-native speech, supporting [16]. Second, other hesitation phenomena – elongations and incomplete turns – are also increased in non-native speech. As a consequence of the above, the time spent producing speech (as opposed to other types of turn construction unit) is longer in native speech. Finally, normalised speech rate is significantly reduced when changing from native to non-native speech, echoing [18].

6. Conclusions and outlook

The current study presents an analysis of L2 speech through the use of a paradigm which allows L1-specific factors to be disentangled from the factor of speaking non-natively. By focusing on easily-measurable acoustic as opposed to phonetic features, we believe the approach can be extended straightforwardly to other language pairs and has the potential to signal nativeness in applications where the style of speech output can be modified to take account of perceived speaker/listener competence in the target language. The DIAPIX-FL Corpus is available as an open resource for further characterisation of speaking in an L2, from segmental factors such as vowel space differences to higher-level interactional and conversational analyses.

Acknowledgements. We thank Valerie Hazan for advice and DiapixUK materials and Julian Villegas, Michael Hobart and Gabriela Cavanagh for their help in annotating sections of the corpus. The research leading to these results was partly funded by the European Community's Future and Emerging Technologies (FET) programme under FET-Open grant number 256230 (LISTA) and by the Spanish MINECO grant FFI2012-31597 (DIACEX).

7. References

[1] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[2] P. K. Kuhl, J. E. Andruski, I. A. Chistovich, L. A. Chistovich, E. V. Kozhevnikova, V. L. Ryskina, E. I. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.

[3] C. E. Snow, "Mothers' speech to children learning language," *Child Development*, vol. 43, no. 2, pp. 549–565, 1972.

[4] M. Uther, M. A. Knoll, and D. Burnham, "Do you speak E-NG-L-I-SH? a comparison of foreigner- and infant-directed speech," *Speech Communication*, vol. 49, no. 1, pp. 2–7, 2007.

[5] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[6] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: a review of human and algorithmic context-induced modifications of speech," *Computer Speech and Language*, vol. 28, pp. 543–571, 2014.

[7] B. Langner and A. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 265–268.

[8] M. Nicolao, J. Latorre, and R. Moore, "C2H A computational model of H&H-based phonetic contrast in synthetic speech," in *Proc. Interspeech*, Portland, USA, 2012.

[9] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise," in *Proc. Interspeech*, Portland, USA, 2012.

[10] C. Valentini-Botinhao, M. Wester, J. Yamagishi, and S. King, "Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise," in *Proc. 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, 2013, pp. 133–138.

[11] B. Langner and A. W. Black, "Using speech in noise to improve understandability for elderly listeners," in *Proc. ASRU*, Cancún, Mexico, 2005, pp. 392–396.

[12] A. Raux, B. Langner, A. W. Black, and M. Eskenazi, "Lets go: Improving spoken dialog systems for the elderly and non-native," in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

[13] O. Kang, D. Rubin, and L. Pickering, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English," *The Modern Language Journal*, vol. 94, no. 4, pp. 554–566, 2010.

[14] M. J. Munro and T. M. Derwing, "The effects of speaking rate on listener evaluations of native and foreign-accented speech," *Language Learning*, vol. 48, no. 2, pp. 159–182, 1998.

[15] S. G. Guion, J. E. Flege, S. H. Liu, and G. H. Yeni-Komshian, "Age of learning effects on the duration of sentences produced in a second language," *Applied Psycholinguistics*, vol. 21, no. 2, pp. 205–228, 2000.

[16] A. Riazantseva, "Second language proficiency and pausing: A study of Russian speakers of English," *Studies in Second Language Acquisition*, vol. 23, pp. 497–526, 2001.

[17] M. J. Munro, T. M. Derwing, and C. S. Burgess, "Detection of nonnative speaker status from content-masked speech," *Speech Communication*, vol. 52, no. 7, pp. 626–637, 2010.

[18] R. L. Rose, "Crosslinguistic corpus of hesitation phenomena: A corpus for investigating first and second language speech performance," in *Proc. Interspeech*, Lyon, France, 2013, pp. 992–996.

[19] T. M. Derwing, M. J. Munro, R. I. Thomson, and M. J. Rossiter, "The relationship between L1 fluency and L2 fluency development," *Studies in Second Language Acquisition*, vol. 31, no. 4, pp. 533–557, 2009.

[20] R. Baker and V. Hazan, "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, 2011.

[21] Council of Europe, *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001. [Online]. Available: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[23] R. M. Uchanski, "Clear speech," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Blackwell, 2008, pp. 207–235.

[24] S. Grawunder and B. Winter, "Acoustic correlates of politeness: prosodic and voice quality measures in polite and informal speech of Korean and German speakers," in *Proc. 5th International Conference on Speech Prosody*, Chicago, USA, 2010.

[25] J. Vaissière, "Perception of intonation," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Blackwell, 2008, pp. 236–263.

[26] D. Bolinger, *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.