Journal of Phonetics 41 (2013) 297–306

Contents lists available at SciVerse ScienceDirect

Journal of Phonetics

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

The role of intonation in language and dialect discrimination by adults

Chad Vicenik *, Megha Sundara

Department of Linguistics, University of California, Los Angeles, 3125 Campbell Hall, Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Received 18 October 2011
Received in revised form
12 February 2013
Accepted 22 March 2013
Available online 5 June 2013

ABSTRACT

It has been widely shown that adults are capable of using only prosodic cues to discriminate between languages. Previous research has focused largely on how one aspect of prosody – rhythmic timing differences – support language discrimination. In this paper, we examined whether listeners attend to pitch cues for language discrimination. First, we acoustically analyzed American English and German, and American and Australian English to demonstrate that these pairs are distinguishable using either rhythmic timing or pitch information alone. Then, American English listeners' ability to discriminate prosodically-similar languages was examined using (1) low-pass filtered, (2) monotone re-synthesized speech, containing only rhythmic timing information, and (3) re-synthesized intonation-only speech. Results showed that listeners are capable of using only pitch cues to discriminate between American English and German. Additionally, although listeners are unable to use pitch cues alone to discriminate between American and Australian English, their classification of the two dialects is improved by the addition of pitch cues to rhythmic timing cues. Thus, the role of intonation cannot be ignored as a possible cue to language discrimination.

1. Introduction

The human ability to distinguish between different languages can provide a window for researchers to explore how speech is processed. After hearing only a very small amount of speech, people can accurately identify it as their native language or not, and if not, can often make reasonable guesses about its identity (Muthusamy, Barnard, & Cole, 1994). This ability to discriminate between languages appears very early in life (Christophe & Morton, 1998; Dehaene-Lambertz & Houston, 1997; Nazzi, Jusczyk, & Johnson, 2000), in some cases, even as early as birth (Mehler et al., 1988; Moon, Cooper, & Fifer, 1993; Nazzi, Bertoncini, & Mehler, 1998). With the use of low-pass filtering and other methods which degrade or remove segmental information, researchers have confirmed that early in acquisition, infants use prosodic cues to distinguish languages (Bosch & Sebastián-Gallés, 1997; Mehler et al., 1988; Nazzi et al., 1998). This reliance on prosodic information for discriminating and identifying languages and dialects continues through adulthood (Barkat, Ohala, & Pellegrino, 1999; Bush, 1967; Komatsu, Mori, Arai, Aoyagi, & Muhahara, 2002; Maidment, 1976, 1983; Moftah & Roach, 1988; Navrátil, 2001; Ohala & Gilbert, 1979; Richardson, 1973).

Although previous research highlights the importance of prosody and its use by human listeners in language identification and discrimination, it remains unclear which sources of prosodic information people use, and if multiple sources are used, how they are integrated with one another. In this paper, we report on a series of acoustic and perceptual experiments to address these questions.

Prosody is a cover term referring to several properties of language, including its linguistic rhythm and intonational system. Languages have frequently been described in terms of their rhythm, since Pike (1946) and Abercrombie (1967), as either “stress-timed” or “syllable-timed,” or if a more continuous classification scheme is assumed, somewhere in between. Evidence suggests that membership in these classes affects the way a language is processed by its native speakers—namely that listeners segment speech based on the rhythmic unit of their language (Cutler, Mehler, Norris, & Segui, 1986; Cutler & Otake, 1994; Mehler, Dommergues, Frauenfelder, & Segui 1981; Murty, Otake, & Cutler, 2007). It has also been suggested that differences in rhythm drive language discrimination by infants (Nazzi et al., 1998, 2000).

Initially, this classification was based on the idea of isochrony. However, research seeking to prove this isochrony in production data has not been fruitful (see Beckman, (1992) and Kohler (2009) for a review). Other researchers have suggested that language rhythm arises from phonological properties of a language, such as the phonotactic permissiveness of consonant clusters, the presence or absence of contrastive vowel length and vowel reduction (Dauer, 1983). This line of thought has led to the development of a variety of rhythm metrics intended to categorize languages into rhythmic classes using measurements made on the duration of segmental intervals (Dellwo, 2006; Grabe & Low, 2002; Ramus, Nespor, & Mehler, 1999; Wagner & Dellwo, 2004; White & Mattys, 2007). Although these metrics have been shown to successfully differentiate between prototypical languages from different rhythm classes on controlled speech materials, they are less successful with uncontrolled materials and non-prototypical

* Corresponding author.
E-mail address: cvicenik@gmail.com (C. Vicenik).

0095-4470/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.03.003

languages, and are not robust to inter-speaker variability (Arvaniti, 2009, 2012; Loukina, Kochanski, Rosner, Keane, & Shih, 2011; Ramus, 2002a; Wiget et al., 2010). Throughout the rest of this paper, when we talk about rhythmic timing information, we are referring to the segmental durational information of the sort captured by these various rhythm metrics.

Despite the limitations of rhythm metrics in classifying languages into rhythmic groups, adult listeners have been shown to be sensitive to the durational and timing differences captured by rhythm metrics when discriminating languages. Ramus & Mehler (1999) re-synthesized sentences of English and Japanese by replacing all consonants with /s/ and all vowels with /a/, and removing all pitch information, forming *flat sasasa* speech. They found that French-speaking adults could discriminate between the two languages (Percent correct: 68%; A′-score: 0.72), indicating that the rhythmic timing information captured by the various metrics does play a role in speech perception—at least when discriminating between rhythmically dissimilar languages. Additionally, there is an evidence that infants rely on rhythmic timing to discriminate some language pairs (Ramus, 2002b). In fact, some researchers predict that infants might use rhythmic timing differences even to distinguish rhythmically similar languages (Nazzi et al., 2000). Thus, the ability to use rhythmic cues to distinguish languages is possible in the absence of experience with either language, and seems to be a language general ability.

Intonation is a second component of prosody that listeners may exploit when discriminating languages. All languages seem to make some use of intonation, or pitch. Pitch is heavily connected with stress in languages that have stress. For example, English often marks the stressed syllable with a specific pitch contour, most commonly a high pitch (Ananthakrishnan & Narayanan, 2008; Dainora, 2001). Pitch contours over the whole sentence consist of interpolated pitch between stressed syllables and phrase-final boundary contours. Languages with weak or no lexical stress still use pitch in systematic ways, often by marking word edges, as in Korean or French (Jun, 2005a; Jun & Fougeron, 2000), making it a universally important component of the speech signal.

Compared to rhythmic timing, listeners' sensitivity to pitch cues when discriminating languages has received little attention. In a pilot study, Komatsu, Arai, and Suguwara (2004) synthesized sets of stimuli, using pulse trains and white noise, to contain different combinations of three cues: fundamental frequency (f0), intensity, and Harmonics-to-Noise Ratio (HNR). All but one of their stimulus conditions contained a re-synthesized amplitude curve matching the original stimuli, from which rhythmic information can potentially be derived. The stimulus condition that had no rhythmic timing information contained only pitch information. They synthesized stimuli corresponding to four languages, English, Spanish, Mandarin and Japanese, which differ both rhythmically and intonationally. Rhythmically, English is considered stress-timed, Spanish syllable-timed and Japanese mora-timed (Ramus et al., 1999). The classification of Mandarin is unclear; it has been described as either stress-timed (Komatsu et al., 2004) or syllable-timed (Grabe & Low, 2002). Intonationally, English and Spanish are both stress (i.e., post-lexical pitch accent) languages, Japanese is a lexical pitch accent language, and Mandarin is a tone language (Jun, 2005b). Averaged across languages, discrimination was possible in all conditions. Discrimination was around 62% when either the rhythmic timing information (the stimuli using various combinations of intensity and HNR) or pitch alone was available. Perhaps unsurprisingly, when both rhythmic timing and pitch cues were available, discrimination was much better, between 75% and 79%.

Other studies have suggested that pitch-based discrimination is possible, even for prosodically-similar languages like English and Dutch (de Pijper, 1983; Willems, 1982), or Quebec and European French (Ménard, Ouellon, & Dolbec, 1999), though in these studies, no effort was made to completely isolate pitch cues from other segmental or prosodic information.

Direct evidence for the role of pitch cues in language discrimination by adults comes from two studies. Using re-synthesized sentences of English and Japanese that had only intonational cues, no segmental or rhythmic information—called *aaaa speech*, Ramus and Mehler (1999) found evidence of discrimination by American English speakers (A′-score: 0.61) but not French speakers. Utilizing the same method of re-synthesis as Ramus & Mehler (1999), Szakay (2008) found that Maori listeners could distinguish between the accents of two New Zealand ethnic groups, Maori English and Pakeha English, at 56% accuracy. Pakeha speakers, on the other hand, were incapable of distinguishing the dialects using only pitch cues. Thus, unlike rhythm, the use of intonation to distinguish languages appears to require experience with at least one of the languages. In addition, depending on the language background of the listener, pitch may not be enough to cue discrimination between languages. Pitch, therefore, may not be as salient a cue as rhythm. Still, pitch is likely as important to speech processing and language discrimination as rhythmic timing properties. Indeed, there is some evidence that pitch may be necessary for infants in language discrimination tasks (Ramus, 2002b).

1.1. Aims of the current study

In this study, we tested whether American English-speaking adults could discriminate their native language and a prosodically-similar non-native language, German, as well as a non-native dialect, Australian English, when segmental information is unavailable. Our goal was to determine what types of prosodic information were necessary to support language discrimination. Specifically, is just pitch information sufficient? Or, do listeners require additional cues, like the rhythmic timing alternation between segments, as captured by the various rhythm metrics to discriminate prosodically-similar languages?

English and German are historically closely related languages. They are rhythmically similar, both are considered stressed-timed languages, and they are intonationally similar, both have tonal phonologies with similar inventories, both tend to position stress word-initially, and both tend to mark stress with a high pitch. These similarities also hold for American English and Australian English, two dialects of the same language.

As stimuli for these experiments, we recorded several hundred sentences in American and Australian English, and German, described below. In Section 2, we examine these recordings acoustically to determine how American English differs from Australian English, and from German, in rhythmic timing and intonation. In Section 3, we describe perception experiments designed to determine whether it is possible to discriminate between prosodically-similar languages/dialects using only prosodic cues, and which cues are necessary and sufficient for adult native English speakers to discriminate these language/dialect pairs.

2. Experiment 1: Acoustic-prosodic measures that distinguish between languages

To determine what types of prosodic information American English-speaking adults could potentially use to discriminate their native language from a prosodically-similar non-native language, and from a non-native dialect of their native language, we acoustically analyzed American, Australian English, and German sentences on two prosodic dimensions—rhythmic timing and pitch, using stepwise logistic regression.

Table 1
Average number of syllables per sentence, average sentence duration, average rate of speech, and minimum, maximum and mean pitch (with standard deviations) for the sentences analyzed in Experiment 1.

	American English	Australian English	German
Average number of syllables/sentence	18 (2)	18 (2)	18 (2)
Average sentence duration (s)	2.95 (0.38)	3.54 (0.52)	3.40 (0.59)
Average rate (syllables /s)	6.10 (0.59)	5.12 (0.63)	5.47 (0.82)
Average minimum pitch (Hz)	117 (40)	127 (48)	115 (29)
Average maximum pitch (Hz)	320 (46)	303 (52)	359 (73)
Mean pitch (Hz)	212 (19)	209 (29)	195 (19)

2.1. Materials

39 English sentences from Nazzi et al. (1998) were recorded by eight female speakers of American English and eight female speakers of Australian English, then translated and recorded by eight female speakers of German in a sound-attenuated booth or quiet room at a sampling rate of 22,050 Hz. Speakers were instructed to read the sentences at a comfortable speaking rate as though to another adult. All American English speakers were from California; all Australian English speakers were from around Sydney; 6 of the 8 German speakers spoke the central German dialect, one spoke upper German, whereas another spoke lower German. Sentences had comparable number of syllables, overall durations, speaking rates, minimum, maximum as well as average pitch as shown in Table 1. 20 sentences from each speaker were selected to form the final stimulus set, with an effort to select for a lack of disfluencies and mispronunciations. These sentences formed a database of 160 sentences per language/dialect. Sentences in the database were also equalized for average intensity at 70 dB using the Scale Intensity function in Praat (Boersma & Weenik, 2006).

2.2. Acoustic measures

2.2.1. Rhythmic measures

As mentioned in Section 1, many metrics have been developed in an attempt to quantify the rhythmic timing of languages (Grabe & Low, 2002; Ramus et al., 1999; Wagner & Dellwo, 2004; White & Mattys, 2007). All metrics have been shown to have strengths and weaknesses (Arvaniti, 2009; Grabe & Low, 2002; Ramus, 2002a), and there has not been any conclusive perceptual research identifying which metric best represents what the listeners attend to. Because we were interested in determining if at all language and dialect pairs could be distinguished using rhythmic information alone, rather than choose between them, we applied all available metrics to our data.

Rhythm metrics traditionally measure intervals of vowel and consonant segments. However, this division can be problematic, particularly in Germanic languages where sonorant consonants often serve as syllabic nuclei. For example, such a division labels the middle syllable in ‘didn't hear’ as part of a single consonantal interval, due to the fully syllabic /n/. Fortunately, the division into vowel and consonant intervals does not appear to be necessary for these metrics to be useful. When based on other divisions, such as voiced and unvoiced segments (Dellwo, Fourcin, & Abberton, 2007) or sonorant and obstruent segments (Galves, Garcia, Duarte, & Galves 2002), rhythm metrics have still been shown to be successfully descriptive. For our data, we segmented and labeled intervals of sonorants and obstruents. As will become clear in the next section, we chose this division primarily for the purposes of re-synthesis because sonorant segments are the segments that carry pitch information, while obstruents obscure pitch.

We used eleven measures of rhythmic timing. For each sentence, we measured the mean percent sonorant interval duration (%S) and the standard deviation of both the obstruent intervals (ΔO) and sonorant intervals (ΔS), analogous to the measures from Ramus et al. (1999), as well as versions of the deviation values corrected for speech rate, VarcoS and VarcoO (Dellwo, 2006; White & Mattys, 2007). The Varco measures require the mean duration of both sonorant and obstruent intervals, which we also included as independent variables in the analysis. Finally, we also measured the raw and normalized pairwise variability index (PVI) values (rPVI and nPVI respectively) for both sonorant and obstruent intervals, analogous to Grabe and Low (2002).

2.2.2. Intonational measures

Unlike rhythm metrics, there are no established metrics for qualifying intonational differences between languages. To operationalize intonation differences, for sonorant segments of each sentence, the only segments that carry pitch, we measured the minimum, maximum and mean pitch (see Baken & Orlikoff (2000) for review), using Praat. We also included the number of pitch rises in each sentence, the average rise height, and the average slope. Pitch rises were identified automatically using a Praat script, and were defined as any minima followed by the closest maxima (i.e., localized) that was greater than 10 Hz: for this purpose any voiceless intervals were ignored.

We focused on pitch rises because all our sentences were declarative. In both dialects of English, and in German, stressed syllables in declarative sentences are typically marked with a high tone preceded by either a shallow rise or by a steep rise (Beckman & Pierrehumbert, 1986; Grice, Baumann, & Benzmuller, 2005; Pierrehumbert, 1980).[1] By counting the number of rises, we expected to be able to capture differences between languages in the frequency of these pitch accents. Measures of the slope were expected to capture differences in pitch accent selection. A language that uses the shallow pitch rise frequently should have a lower average slope than languages which use a steeper rise more frequently.

[1] In the ToBI-transcription system, a shallow rise to a high tone would be labeled as a H^*, and a steep rise would be labeled as a $L+H^*$, or occasionally a L^*+H.

2.3. Results and discussion

In Table 2, we present the means and standard deviations for each rhythm and intonation measure for American English, Australian English and German. Whether or not listeners are specifically using the information captured by the various rhythm or intonation metrics, there are significant differences between the two language pairs in both rhythmic timing and pitch measures as compared using *t*-tests.

To test these differences further, we conducted a stepwise, binary logistic regression for each language pair in order to see how much of the data could be correctly classified using these measures. American English was separately compared to German and to Australian English. We used logistic regression as an alternative to discriminant analysis because it requires fewer assumptions. Namely, logistic regression does not require independent variables to be normally distributed or have equal within-group variances. First, the 11 rhythm measures described above were used as independent variables. Classification scores are reported in Fig. 1.

Overall, using rhythm measures alone, the model was able to accurately classify the two pairs over 70% of the time. This is well above chance, and somewhat surprising, considering the three tested languages are all stressed timed, and so expected to be rhythmically very similar. However, no single rhythmic timing measure or set of measures generated this high classification accuracy. The top two independent variables that were included in each model—the percentage of the sentence that was sonorant (%S) and the nPVI index for sonorants for American English vs. German, and the mean obstruent duration (MeanO) and the nPVI index for obstruents for American vs. Australian English—were different. Thus, it is likely that the model is exceptionally good at taking advantage of the very fine differences present in the data.

We also ran logistic regressions testing the classification when only pitch measures were used as predictors. These results are also presented in Fig. 1. Overall, using pitch cues alone, the logistic regression model was able to correctly classify about 80% of the sentences. Although the three languages are similar in their tonal inventories, we take the high classification rates based on the pitch cues alone as support for the existence of differences in how pitch is employed by the different languages.

Table 2
Means and standard deviations for each rhythm and pitch measure for the sentence stimulus set in American English, Australian English and German. *T*-test comparisons between American and Australian English, and American English and German for each measure are also presented.

	American English	Australian English	American vs. Australian English	German	American English vs. German
Sentence duration	2.95s (0.38)	3.54s (0.52)	$t(318)=11.313$, $p<0.001$	340s (0.59)	$t(318)=7.994$, $p<0.001$
Speech rate	6.10 syl/s (0.59)	5.12 syl/s (0.63)	$t(318)=14.187$, $p<0.001$	5.47 syl/s (0.82)	$t(318)=7.777$, $p<0.001$
%Son	59.51% (5.61)	58.72% (5.91)	$t(318)=1.229$, $p=$n.s.	54.72% (6.76)	$t(318)=6.900$, $p<0.001$
sd Son	93.16 (30.04)	110.19 (36.02)	$t(318)=4.592$, $p<0.001$	77.73 (35.37)	$t(318)=4.208$, $p<0.001$
sd Obs	49.58 (15.82)	59.98 (18.92)	$t(318)=5.334$, $p<0.001$	58.07 (16.57)	$t(318)=4.690$, $p<0.001$
rPVI Obs	57.46 (17.34)	66.73 (23.39)	$t(318)=4.027$, $p<0.001$	66.37 (19.11)	$t(318)=4.370$, $p<0.001$
nPVI Obs	65.35 (15.70)	59.68 (15.82)	$t(318)=3.219$, $p=0.001$	66.25 (15.53)	$t(318)=0.514$, $p=$n.s.
Mean Obs	93.43 (14.72)	113.23 (18.55)	$t(318)=10.577$, $p<0.001$	102.74 (18.65)	$t(318)=4.955$, $p<0.001$
rPVI Son	104.31 (37.59)	124.72 (45.56)	$t(318)=4.369$, $p<0.001$	85.03 (39.56)	$t(318)=4.471$, $p<0.001$
nPVI Son	71.74 (16.43)	73.55 (17.19)	$t(318)=0.961$, $p=$n.s.	62.15 (13.68)	$t(318)=5.674$, $p<0.001$
Mean Son	141.17 (28.94)	165.87 (37.26)	$t(318)=6.624$, $p<0.001$	128.65 (40.02)	$t(318)=3.206$, $p=0.001$
Varco Obs	52.90 (12.71)	52.74 (13.20)	$t(318)=0.106$, $p=$n.s.	56.81 (12.51)	$t(318)=2.777$, $p<0.006$
Varco Son	65.52 (13.50)	65.96 (12.32)	$t(318)=0.308$, $p=$n.s.	59.10 (12.51)	$t(318)=4.415$, $p<0.001$
Min F0	117.35 (39.95)	126.97 (48.35)	$t(318)=1.941$, $p=$n.s.	114.6 (29.26)	$t(318)=0.701$, $p=$n.s.
Max F0	320.35 (46.33)	303.12 (51.72)	$t(318)=3.138$, $p=0.002$	358.9 (73.30)	$t(318)=5.629$, $p<0.001$
Mean F0	211.97 (18.72)	208.62 (29.41)	$t(318)=1.216$, $p=$n.s.	195.03 (19.39)	$t(318)=7.949$, $p<0.001$
Number of rises	7.52 (2.47)	10.55 (2.70)	$t(318)=10.498$, $p<0.001$	9.18 (2.64)	$t(318)=5.828$, $p<0.001$
Average rise (F0)	39.41 (12.62)	36.24 (11.14)	$t(318)=2.382$, $p=0.018$	55.35 (23.02)	$t(318)=7.682$, $p<0.001$
Average slope	506.5 (493.30)	491.49 (434.80)	$t(318)=0.288$, $p=$n.s.	1137.24 (1384.41)	$t(318)=5.429$, $p<0.001$

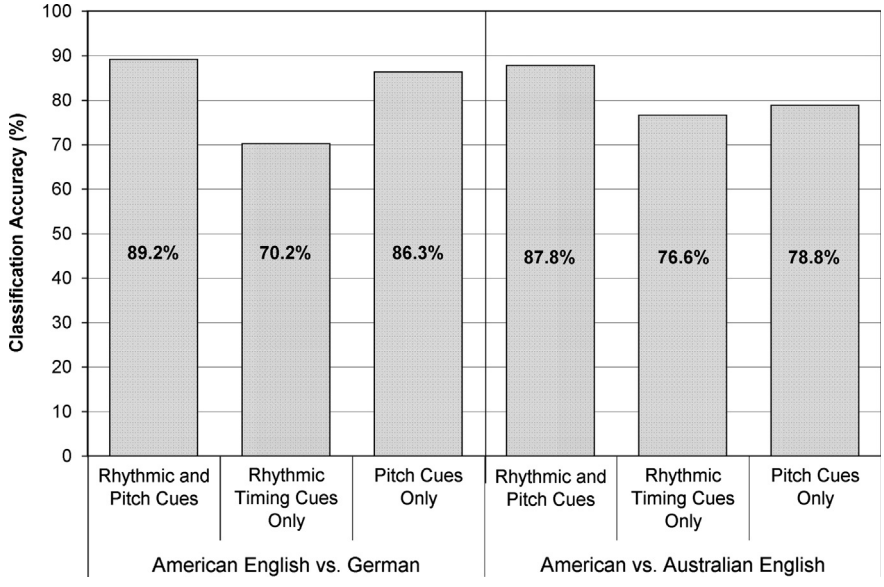

Fig. 1. Classification scores from a logistic regression for the two language/dialect pairs under different conditions: the combination of rhythmic timing and pitch information, rhythmic timing information only, and pitch information only.

A comparison of classification accuracy using rhythmic timing and pitch information gave a different hierarchy of usefulness of cues for each language/dialect pair. For American English vs. German, classification was higher using pitch measures when compared to rhythmic timing measures ($\chi^2(1)=7.61$, $p=0.006$). However, for American vs. Australian English, classification using pitch cues alone was comparable to classification using rhythmic timing cues alone ($\chi^2(1)=0.35$, $p=0.554$.).

Finally, although classification using just rhythmic timing measures or just pitch measures was high, it was expected that when the regression model had access to both types of measures, it would perform even better. Classification using the combined cues was significantly better than using only rhythmic timing cues ($\chi^2(1)=11.16$, $p<0.001$), but there was little improvement in the classification of American English and German when rhythm measures were added to the model in addition to the pitch information ($\chi^2(1)=0.39$, $p=0.532$).

In contrast, the classification of American and Australian English was better with both cues when compared to each cue alone. There was a significant improvement with access to both cues than with rhythmic timing cues alone ($\chi^2(1)=4.29$, $p=0.038$). Model performance was also marginally better with both cues compared to performance with pitch cues alone ($\chi^2(1)=2.91$, $p=0.088$).

In summary, logistic regression using acoustic measures of rhythmic timing and pitch showed that both language pairs could be classified using either cue type. Based on these acoustic differences, adult listeners should be able to discriminate between either American English and German, and American and Australian English using rhythmic timing or pitch cues. Further, for American English vs. German, we expect pitch cues to be more informative than rhythmic timing cues. For American and Australian English, we expect the combination of rhythmic timing and pitch to be more informative than either cue alone.

3. Experiment 2: The perceptual role of rhythmic timing and pitch

The previous section showed that the two language pairs, American and Australian English, and American English and German, can be distinguished acoustically using only rhythmic timing or pitch information. This section explores whether adult listeners can use prosodic information alone to distinguish American English and German, and American and Australian English, and if so, which cues they use. To do this, we conducted three perceptual experiments, each testing a different combination of prosodic cues.

The first, *filtered,* condition tested discrimination using low-pass filtered speech (filtered above 400 Hz). Low-pass filtering is a method that removes segmental information from speech, but leaves behind prosodic information—including rhythmic timing and pitch.

In the second experimental condition, we tested discrimination using re-synthesized speech similar to the *flat sasasa* speech used in Ramus and Mehler (1999). This method of re-synthesis completely removes segmental information and pitch information, as well as any other prosodic information, leaving only rhythmic timing information intact. The only cues available to listeners in this condition are rhythmic timing of sonorant and obstruent segments. We refer to this condition as the *rhythmic timing only* condition.

Finally, in the third condition, the intonational contours from the original stimuli were re—synthesized onto a long, continuous /a/ sound, forming an *intonation only* condition. Breaks in the original pitch contour caused by obstruent sounds were replaced with interpolation. This was done to obscure rhythmic timing information.

Based on the acoustic analyses presented in the first experiment, there is enough information contained in the pitch contours and rhythmic timing patterns of the experimental stimuli to successfully discriminate between the languages. Thus, it is possible that listeners may discriminate in all conditions. However, we suspect the regression models are performing above human capabilities, so it should not be surprising if listeners fail in one or more conditions.

For American English vs. German, the models' classification scores using only pitch cues were as high as when using both rhythmic timing and pitch, and significantly better than when using only rhythmic timing cues. If listener's follow this same pattern, we predict listeners should be as good at discriminating in the *intonation only* condition as they are in the *filtered* condition, when both pitch and rhythmic timing cues are available. Furthermore, they should perform better in both these conditions than in the *rhythmic timing only* condition.

For American vs. Australian English, the models' classification scores using both rhythmic timing and pitch were significantly higher than when using either cue type alone. Thus, we predict listeners should be better at discriminating in the *filtered* condition than in either the *rhythmic timing* or *intonation only* conditions.

3.1. Methods

3.1.1. Stimuli

All 480 sentences from Experiment 1 were used for each condition (160 for each of the three languages). In the *filtered* condition, sentences were low-pass filtered using Praat at a frequency cut-off of 400 Hz, with 50 Hz smoothing.

In the *rhythmic timing only* condition, the sentences were re-synthesized. Sonorant segments were replaced with /a/ and obstruent segments were replaced with silence, simulating a glottal stop, producing new sound files of the same length as the original. This method of re-synthesis was simpler to set up, and avoids any issues with co-articulation, or its absence, between the sonorant and obstruent segments because glottal stop and /a/ show no formant transitions. A similar form of re-synthesis was used in Szakay (2008), though with consonants and vowels rather than obstruents and sonorants. The original set of sentences had no disfluencies or sentence-medial pauses; therefore, all periods of silence in the re-synthesized stimuli corresponded to obstruent segments. However, there was no differentiation between sentence onset and offset consonants, and surrounding silence. Thus, information about any obstruents located on the sentence edges was lost in the re-synthesis. A new logistic regression model showed that classification scores for the languages were similar to those found in Section 2 (76.5% for American English and German, which is, surprisingly, a slight but non-significant improvement on the acoustic analysis previously reported in Section 2; originally, 70.2%; χ^2 (1)=1.02, $p=0.313$; and 76.6% for American and Australian English, identical to the score reported in Section 2). Nor did this change alter which rhythmic measures, shown in Table 2, the language pairs significantly differed on. In the final step of re-synthesis, sentences were given a flat, monotone pitch contour of 200 Hz, which is near the mean pitch across all sentences (205 Hz). The resulting sentences, thus, only contained rhythmic timing information.[2]

[2] It is possible that listeners were not treating the re-synthesized stimuli in this experiment in the same way as normal speech, adding a potential confound to the study. For example, listeners might be treating the silences as pauses, rather than obstruent (or consonantal) intervals. This seems unlikely due to the small duration of the silent intervals (mean=103 ms;

Finally, for the *intonation only* condition, the pitch contours of the original sentences were re-synthesized onto a long, continuous /a/ vowel using Praat. The length of the base vowel matched the duration of the original sentence. Pitch was interpolated over obstruent portions of the original sentence, forming a single, continuous contour. This is the same method of re-synthesis used by Ramus and Mehler (1999) and Szakay (2008). It should be noted that this method of re-synthesis adds some information to the signal – namely, the interpolated pitch. However, this condition was intended to test discrimination using only pitch. Had the intervals of silence been preserved, both pitch and rhythmic timing cues would be present.

3.1.2. Participants

98 native speakers of American English were recruited from the undergraduate population of UCLA. Most received course extra credit, some were paid. Participants who spoke either German or Australian English, or had ever traveled to either country were excluded ($n=8$), thus, 90 participants were included in the final analysis.

3.1.3. Procedure

30 subjects were used in each of the conditions, and for each condition, subjects were divided into two groups. 15 subjects heard American English vs. German and 15 heard American English vs. Australian English.

The experiments were presented in Praat in a sound-attenuated booth. Sentences were presented one at a time over loudspeakers at an average intensity of 70 dB. Each participant heard 320 sentences, 160 in each of the two languages/dialects. Sentences were presented in a randomized order. Subjects were told they would hear a number of sentences and had to decide if they were spoken in American English or some other language/dialect. They were not informed about either the number of foreign languages, dialects, or their identity. After each sentence was played, participants identified it as “American English” or “Other.” Testing lasted for around 30 min.

3.2. Analyses

Percent correct scores for each condition are presented in Table 3. To take into account any response bias subjects may have had, we converted the responses into hit-rates and false alarm-rates. Correctly identified American English sentences were counted as hits; sentences misidentified as American English were counted as false alarms. Discrimination scores (A′) were then calculated, and are also presented in Table 3. A′-scores are a non-parametric analog of d′-scores. They range from 0 to 1, where chance performance is 0.5. The higher the A′-score, the more accurate participants were in discriminating the language pairs. Analysis of percent correct data, d′- and A′-scores show identical patterns, thus throughout the paper, only analyses with A′ are reported. A one-sample t-test was conducted to compare the group A′ scores to chance (0.5).

Subjects’ data were also examined individually to see if they scored above chance, in order to determine whether individual performance conforms to group trends. To determine whether a subject performed significantly above chance, the 95% confidence limits were calculated based on the normal approximation to the binomial distribution (Boothroyd, 1984; Kishon-Rabin, Haras, & Bergman, 1997).[3] The confidence limits were calculated based on the number of trials ($n=320$), the probability of guessing (1/2) and the *t*-value (2, if number of trials is more than 20). These parameters hold for all experiments. Subjects with A′-scores above 0.556 were considered to have performed significantly above chance with a $p<0.05$. These results are also presented in Table 3, as well as in Figs. 2 and 3. Finally, the number of subjects performing above chance was compared across conditions using a chi-square test.

3.3. Results and discussion

American English-speaking adults were able to successfully discriminate their native language from a non-native language, German, or from a non-native dialect, Australian English, in a majority of the experimental conditions. However, in all conditions, discrimination was quite difficult, indicated by the proximity of the average accuracy to chance. Across all conditions, the average percent correct score was 53.6% and the average A′-score was 0.56. This is quite a bit lower than the classification rates obtained by the regression models in Experiment 1, which ranged between 70% and 90%, confirming the suspicion that the regression models are outperforming human ability. Listeners simply are not as capable of utilizing the fine prosodic differences between the languages captured by the acoustic analysis. Rather, they likely rely more on segmental information to discriminate in more natural tasks. However, an alternate possibility may be that the measures used in the acoustic analysis do not accurately capture the information perceptually extracted by human listeners.

Listeners were able to successfully discriminate between American English and German in all three experimental conditions. Thus, the rhythmic timing information and the pitch information available in the stimuli were each sufficient to allow for discrimination between the two languages. Based on the results of the acoustic analysis in Experiment 1, it was expected that discrimination would be easier in both the *filtered* and *intonation only* conditions compared to the *rhythmic timing only* condition. A one-way ANOVA was used to compare A′-scores across the three conditions, but found no significant differences ($F(2,42)=0.335$, $p=$n.s.), nor was there any difference in the number of participants performing above chance across the conditions (*rhythmic timing only* vs. *intonation only*: $\chi^2(1)=0$; *rhythmic timing only/intonation only* vs. *filtered*: $\chi^2(1)=0.56$, $p=0.454$). Thus, listeners

(*footnote continued*)

SD=19.3). However, to confirm that this was not the case, we replicated the *rhythm only* American English–German condition using *monotone sasasa* speech, as in Ramus and Mehler (1999). To create this stimuli, sonorant intervals in the original recorded stimuli were replaced with /a/ and obstruent intervals were replaced with /s/. A flat pitch contour of 200 Hz was synthesized onto the resulting stimuli. Fifteen native American English-speaking adults were run on this additional condition. Results are discussed in footnote 3.

[3] Confidence limits for scores expected from guessing depend on the number of trials (n), and on the probability of guessing (p; in a two-alternative forced-choice, $p=\frac{1}{2}$

$$\text{Confidence limit} = \text{Chance score} \pm t*\frac{\sqrt{p(1-p)}}{n}$$

Values for t are taken from the t-tables according to $n-1$ degrees of freedom. For $n=20$ or more, $t=2.0$ for 95% confidence limits. In our experiments, chance is at 0.5, thus

$$95\%\ \text{confidence limit} = 0.5 \pm \sqrt{\frac{1}{n}}$$

Table 3
Results from Experiment 2, reporting group percent correct scores, A′-scores, the number of participants performing above chance and statistical tests comparing A′-scores to chance (0.5). The *filtered* condition tests discrimination using low-pass filtered sentences of American English and German, and American and Australian English. The *rhythm-only* condition tests discrimination using re-synthesized sentences containing only rhythmic timing information. The *intonation only* condition tests discrimination using sentences containing only pitch information.

Experiment condition	Language/dialect pair	Percent correct	A′ score	Participants above chance	Comparison to chance
Filtered	vs. German	53.5	0.56	8/15	$t(14)=3.0$, $p=0.009$
	vs. Australian	54.5	0.58	11/15	$t(14)=4.2$, $p=0.001$
Rhythm Only	vs. German	54.1	0.57	10/15	$t(14)=2.8$, $p=0.015$
	vs. Australian	53.3	0.56	6/15	$t(14)=2.5$, $p=0.028$
Intonation Only	vs. German	55.6	0.59	10/15	$t(14)=3.2$, $p=0.007$
	vs. Australian	50.7	0.50	5/15	$t(14)=0.02$, n.s.

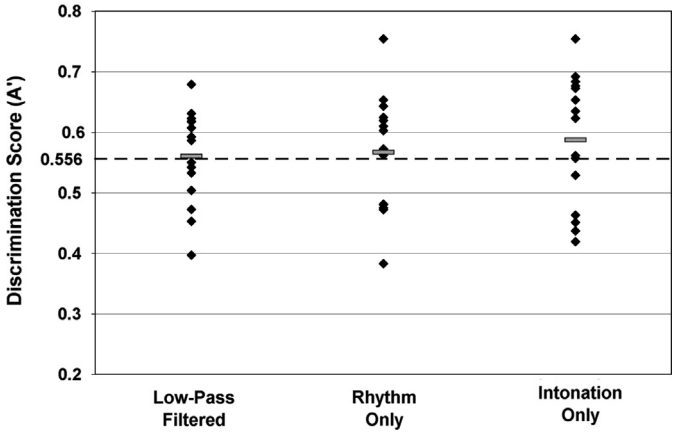

Fig. 2. Results from the American English vs. German conditions of Experiment 2, showing discrimination scores (A′) of each subject (black dots), the group average (gray bar), and the 95% confidence interval line (at 0.556).

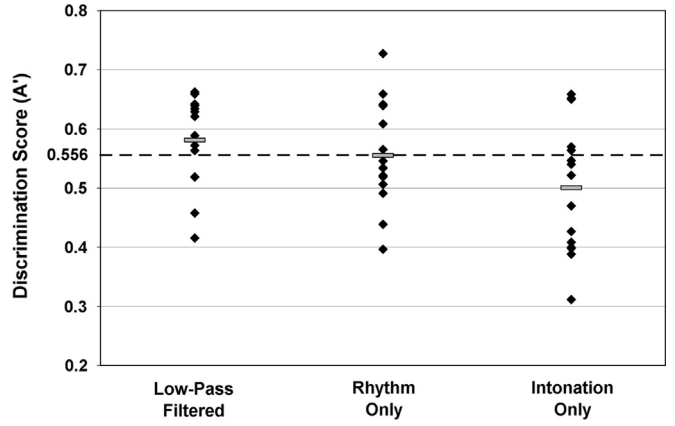

Fig. 3. Results from the American vs. Australian English conditions of Experiment 2, showing discrimination scores (A′) of each subject (black dots), the group average (gray bar), and the 95% confidence interval line (at 0.556).

were equally good at using rhythmic timing information or pitch to discriminate between American English and German. Further, access to both cues did not improve listeners' performance.[4]

For American and Australian English, it was expected that listeners would perform better at discriminating between the dialects when they had access to both rhythm and intonation cues than when they had access to only one set of cues. Listeners were able to successfully discriminate between the two dialects in the *filtered* and *rhythmic timing only* conditions, but not in the *intonation only* condition. A one-way ANOVA comparing A′-scores across the three conditions found marginally significant differences across the conditions ($F(2,42)=3.027$, $p=0.059$). A Tukey's HSD post-hoc test showed that this marginal effect was being driven by the difference in performance on the *filtered* and *intonation only* conditions ($p=0.052$). Matching the difference in group performance, there were significantly more participants who discriminated better than chance in the *filtered* condition than in the *intonation only* condition ($\chi^2(1)=4.82$, $p=0.028$). Consistent with the acoustic analysis previously reported, these results clearly indicate

[4] For the *sasasa speech*, the group average percent correct score was 55% and average A'-score was 0.58, which was significantly greater than chance ($t(14)=4.675$; $p<0.001$). A'-scores were not significantly different from the *rhythm only* experiment. Ten of 15 subjects performed significantly greater than chance, an identical number to the original *rhythmic timing-only* condition. Thus, listeners treated the intervals of silence as segmental in the same way as intervals of /s/.

that listeners were better at discriminating between the two dialects when they had access to both rhythmic timing and pitch than when they only had access to pitch. In fact, listeners were not capable distinguishing between American and Australian English using only pitch cues.

Further, there was no difference in A′-scores between the *rhythmic timing only* and *filtered* condition. However, the difference in the number of participants who discriminated significantly better than chance in the two conditions was marginally significant ($\chi^2(1)=3.39$, $p=0.065$). Thus, as expected, even though listeners could still discriminate American and Australian English using only rhythmic information, they were better when they had access to both rhythmic timing and pitch cues.

4. General discussion

In this study, we examined whether it was possible to discriminate between closely related languages – American English and German, and American and Australian English – using prosodic cues alone. Using a logistic regression analysis, we showed that the two language pairs were acoustically distinct, in both rhythmic timing and pitch. Classification accuracy was lowest with rhythmic timing cues alone, but still well above chance. Classification between American English and German was significantly better using only pitch cues than only rhythm, and classification accuracy did not improve further when the model included rhythmic timing cues in addition to pitch cues. For American and Australian English, there was no difference in classification accuracy when only rhythmic timing or only pitch cues were used, but there was significant improvement in classification when the model had access to both cues.

Next, with perception experiments involving low-pass filtered and re-synthesized stimuli, we demonstrated that American English listeners could discriminate between both language/dialect pairs using prosodic cues alone, though with difficulty. When segmental information had been stripped from the stimuli, successful discrimination was only slightly better than chance. When above chance, listeners in our experiments scored around 54% correct (A′-score of around 0.57) for both language pairs. Our results are comparable to previous discrimination studies on prosodically similar languages using segmentally degraded stimuli. For example, Barkat et al. (1999) found that listeners in their study, using only prosodic cues, correctly distinguished between dialects of Arabic 58% of the time. Similarly, Szakay (2008), using stimuli re-synthesized as in the current study, found that listeners could distinguish between New Zealand dialects with an average accuracy of 56%.

The low discrimination scores seen in the current and previous studies indicate how heavily adult listeners rely on segmental information when processing speech. Using unmodified, full-cue speech, listeners are expected to perform near ceiling, especially in the cases where they are discriminating between their native language and a non-native language. It is also clear that, although listeners may be able to use prosodic cues alone to discriminate between languages, they are not very good at it. Acoustically, there is a wealth of information listeners could use to help them classify languages, but listeners only seem capable of utilizing a fraction of it.

Crucially, in this study, we examined whether adult listeners were able to use pitch cues for language discrimination. The results differed for the two language pairs. For American English and German, both rhythmic timing information and sentential pitch information were each sufficient, on their own, to cue discrimination. Further, having rhythmic timing and pitch information together did not improve listeners' performance. For American and Australian English, rhythmic information was sufficient to cue discrimination, though it should be noted that as a group, listeners were just barely significantly better than chance. Pitch information was not sufficient to cue discrimination. However, there was evidence that access to both cues did facilitate and improve listeners' ability to discriminate between the dialects than when they had access to either rhythmic timing or pitch alone.

Typically, rhythm, and the rhythm metrics in particular, are often discussed in terms of their ability to classify languages into different rhythm classes. After all, it is the broad classification that is thought to be important in speech processing because, for example, of the idea that listeners develop speech segmentation strategies around either syllables, feet or morae (Cutler et al., 1986; Cutler & Otake, 1994; Mehler et al., 1981). Rhythm metrics, of course, most directly measure duration and variability in duration of different segments, but are often equated with or treated as an operationalized version of linguistic rhythm. It would be expected, then, that both adults (Ramus & Mehler, 1999) and infants (Ramus, 2002b) can discriminate between languages from different rhythm classes using only segmental duration and timing information.

However, a considerable amount of recent research has cast doubt on the ability of rhythm metrics to accurately make the clear classifications often demanded of them. For example, it has been shown that inter-speaker variability in segmental rhythm can often be as large as or larger than cross-linguistic differences, which makes classification into rhythmic groups very difficult (Arvaniti, 2009; Loukina et al., 2011). It is all the more surprising, then, that listeners in the current study were able to discriminate between languages using only segmental duration and timing information. The languages tested in this study are considered to be rhythmically similar—specifically, they are all considered to be stress-timed languages. The stimuli, at least for American and Australian English, consisted of identical sentences, which would presumably produce very similar segmental rhythm patterns. Yet, the application of the rhythm metrics classified the two language/dialect pairs reasonably well, and listeners were able to discriminate between them at a greater than chance accuracy. Thus, despite the contentious relationship between rhythm metrics and actual linguistic rhythm, the role segmental duration and variability plays in speech processing cannot be discounted.

In the ongoing attempts to properly define linguistic rhythm, several researchers have suggested that the focus on rhythm metrics has distracted from other properties of speech that may influence rhythmic perception, namely pitch (Arvaniti, 2009; Kohler, 2009). It is definitely true that pitch matters in speech processing, including language discrimination. The experiments in the current study show that pitch information alone is enough to allow listeners to distinguish between American English and German. If, indeed, the cognitive processing of rhythm involves the integration of segmental timing and pitch information, as well as possibly other prosodic properties of speech, we might expect listeners to be better at discriminating languages when given access to both cues. This was only partially supported by the listening experiments: the addition of pitch cues to rhythm cues improved listener classification of American vs. Australian English, but not American English vs. German.

Why was there no observed improvement in discrimination between American English and German when listeners had access to both rhythm and pitch cues? It is possible that variability in the prosodic properties of the stimuli in the current experiments hindered listeners' performance, particularly in the American English vs. German case. Unlike the American and Australian English speakers, the German speakers were not tightly controlled for dialect background. However, comparing the percent correct scores for individual speakers (in Table 4) reveals no obvious patterns that can be traced to dialect differences. The identification accuracy for the Upper German speaker was comparable to that of the 6 central German speakers (ranked 4th on *filtered* and *rhythm only* conditions and 5th on the *intonation only* condition). The low German speaker is the most accurately identified speaker with just intonation cues alone and second most accurately identified in the *rhythm only* and *filtered* conditions. However, the overall discrimination patterns do not change if this speaker is removed from the analysis.

Table 4
Listeners' average percent correct scores for each German speaker in each experimental condition. Speakers' dialect and region of origin are also provided.

Speaker	Filtered	Rhythm Only	Intonation Only	Dialect (Broadly)	Region of Origin
1	0.567	0.563	0.550	Central German	Upper Saxon
2	0.572	0.560	0.570	Upper German	Austria
3	0.711	0.650	0.567	Central German	Frankfurt
4	0.578	0.587	0.610	Low German	West Berlin
5	0.528	0.580	0.543	Central German	Bonn
6	0.561	0.540	0.517	Central German	Unknown
7	0.656	0.526	0.583	Central German	Trier
8	0.578	0.533	0.573	Central German	Trier

Another possibility is that, if indeed the integration of different prosodic cues allows the listener a fuller perception of linguistic rhythm, the more information a listener has, the harder it may be to distinguish languages like English and German, which are both stress-timed. If rhythm classes have psychological reality, then it should be harder to discriminate within categories than across categories. However, this does not hold true for American and Australian English, which by most accounts, should be more prosodically similar than American English and German. Listeners were better at discriminating between those dialects when given access to multiple prosodic cues.

Rather, it seems as if listeners simply integrate rhythm and pitch differently for different language pairs. The language specificity of this effect likely has as much to do with the listener's linguistic background as it does the languages they are listening to. For example, Maori listeners could use pitch alone to distinguish Maori and Pakeha English, but Pakeha listeners could not (Szakay, 2008). Similarly, speakers of Midland American English viewed their dialect as more perceptually distant from Northern American English than speakers of the latter dialect did (Clopper, 2007). Therefore, it remains to be determined how German listeners might use rhythm and pitch cues, or their combination, to discriminate between American English and German. Regardless, the difference in the usefulness of the pitch cues across the two pairs, but not the rhythm cues, suggests that these cues are not automatically integrated into one percept. Rather, any such integration (or lack thereof) is language specific. Future studies will be needed to outline the nature of this integration.

Finally, results from the current study have implications for language acquisition. It is well known that infants can discriminate between certain languages as early as birth (Mehler et al., 1988). It has been proposed that this ability is driven by rhythm, that infants are discriminating languages from different rhythm classes (Nazzi et al., 1998). Discrimination of rhythmically-similar languages is acquired at a later age, around 4- to 5-months, and is thought to require a familiarity with at least one of the languages (Bosch & Sebastián-Gallés, 1997; Nazzi et al., 2000). For example, American English-learning infants are able to discriminate between English and Dutch, and American and British English, but not German and Dutch. What drives discrimination for older infants?

Nazzi et al. (2000) suggest that 4- to 5-month-olds use rhythmic information to discriminate rhythmically-similar languages. Specifically, they suggest infants at this age have learned the specific rhythmic properties of their native language, including the properties that distinguish it from other languages of the same rhythm class. Indeed, our results with adult listeners suggest it is possible that infants are discriminating languages using fine differences in segmental duration and variability. However, it is equally possible that infants are relying on differences in pitch to discriminate languages. Additional research is needed to determine if, like adults, infants are capable of using both rhythm and intonational cues to discriminate languages.

5. Conclusion

In this paper, we have shown that adults can discriminate between prosodically-similar languages using either rhythmic timing or pitch. Their performance is low, but consistently above chance. However, not all language pairs can be discriminated using a single cue type. Some pairs are discriminable using either rhythmic timing or pitch, e.g., American English and German, while others are discriminable using rhythmic timing information, but not pitch, e.g., American and Australian English. Nevertheless, both rhythmic timing and pitch play crucial roles in language discrimination by adults.

Acknowledgments

We would like to thank Christina Kitamura, Volker Dellwo, Theirry Nazzi and Constanze Weise for help in recording experimental stimuli and Joseph Randazzo for help in experimental setup and data collection. We would also like to thank Sun-Ah Jun, Patricia Keating, Bruce Hayes, Toby Mintz, the members of the UCLA Phonetics Lab and reviewers for comments. This study is part of C. Vicenik's doctoral thesis. It was supported by a UCLA Summer Research Mentorship Award (2008) to C. Vicenik, a UCLA COR Faculty Research Grant (2008–2009), and an NSF grant (BCS-0951639, 2010–2013) to M. Sundara. Parts of this research were presented at the 2nd Acoustical Society of America workshop on "Cross-language speech perception and variations in linguistic experience," Portland (Abstract in JASA, 125(4), 2764) and at Acoustics'08, the 155th meeting of the Acoustical Society of America, Paris (Abstract in JASA, 123(5), 3883).

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.
Ananthakrishnan, S., & Narayanan, S. (2008). Fine-grained pitch accent and boundary tone labeling with parametric F0 features. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 4545–4548). Las Vegas, Nevada.
Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*, 46–53.
Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*, 351–373.
Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice* (pp. 145–224)San Diego: Singular Publishing Group145–224.
Barkat, M., Ohala, J. J., & Pellegrino, F. (1999). Prosody as a distinctive feature for the discrimination of Arabic dialects. *Proceedings of Eurospeech*, *99*, 395–398.

Beckman, M. E. (1992). Evidence for speech rhythms across languages. In: Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 457–463). Tokyo: OHM Publishing Co.
Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309.
Boersma, P., & Weenik, D. (2006). *Praat: Doing phonetics by computer* (*version* 5.2.22). ⟨http://www.praat.org/⟩.
Boothroyd, A. (1984). Auditory perception of speech contrasts by subjects with sensorineural hearing loss. *Journal of Speech and Hearing Research*, *27*, 134–144.
Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*, 33–69.
Bush, C. (1967). Some acoustic parameters of speech and their relationship to the perception of dialect differences. *TESOL Quarterly*, *1*, 20–30.
Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science*, *1*, 215–219.
Clopper, C. (2007). Free classification of regional dialects of American English. *Journal of Phonetics*, *35*(3), 421–438.
Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.
Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, *33*, 824–844.
Dainora, A. (2001). *An empirically based probabilistic model of intonation in American English*. University of Chicago [dissertation].
Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*, 51–62.
de Pijper, J. R. (1983). *Modeling British English intonation* (pp. 1–152)Dordrecht: Foris1–152.
Dehaene-Lambertz, G., & Houston, D. (1997). Faster orientation latencies toward native language in two-month-old infants. *Language and Speech*, *41*, 21–43.
Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for delta C. In: *Language and language processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003*, (pp. 231–241). P. Karnowski & I. Szigeti [Eds]. Frankfurt am Main, Germany: Peter Lang Publishing Group.
Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In: *Proceedings of 16th international congress of phonetic sciences (ICPhS)* (pp. 1129–1132). Saarbrücken, Germany.
Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. *Speech Prosody*, *2002*, 323–326 [in Aix-en-Provence].
Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In: C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology* (pp. 515–546). Berlin: Mouton de Gruyter.
Grice, M., Baumann, S., & Benzmuller, R. (2005). German intonation in autosegmental-metrical phonology. In: S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 55–83). New York: Oxford University Press.
Jun, S.-A. (2005a). Korean intonational phonology and prosodic transcription. In: S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 201–229). New York: Oxford University Press.
Jun, S.-A. (2005b). Prosodic typology. In: S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 430–458). New York: Oxford University Press.
Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In: A. Botinis (Ed.), *Intonation: Analysis, modeling and technology* (pp. 209–242). Kluwer Academic Publishers.
Kishon-Rabin, L., Haras, N., & Bergman, M. (1997). Multisensory speech perception of young children with profound hearing loss. *Journal of Speech, Language and Hearing Research*, *40*, 1135–1150.
Kohler, K. J. (2009). Rhythm in speech and language: A new research paradigm. *Phonetica*, *66*, 29–45.
Komatsu, M., Arai, T., & Suguwara, T. (2004). Perceptual discrimination of prosodic types and their preliminary acoustic analysis. *Proceedings of Interspeech*, *2004*, 3045–3048.
Komatsu, M., Mori, K., Arai, T., Aoyagi, M., & Muhahara, Y. (2002). Human language identification with reduced segmental information. *Acoustical Science and Technology*, *23*, 143–153.
Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America*, *129* (5), 3258–3270.
Maidment, J. A. (1976). Voice fundamental frequency characteristics as language differentiators. *University College, London, Speech and Hearing: Work in Progress*, *2*, 74–93.
Maidment, J. A. (1983). Language recognition and prosody: Further evidence. *University College, London, Speech, Hearing and Language: Work in Progress*, *1*, 133–141.
Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*, 298–305.
Mehler, J., Jusczyk, P. W., Lambertz, G., Halstead, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*, 143–178.
Ménard, L. Ouellon, C., & Dolbec, J. (1999). Prosodic markers of regional group membership: The case of the French of Quebec versus France. In *Proceedings of the XIVth international congress of phonetic sciences (ICPhS 99)*(pp. 1601–1604). San Francisco.
Moftah, A., & Roach, P. (1988). Language recognition from distorted speech: Comparison of techniques. *Journal of the International Phonetic Association*, *18*, 50–52.
Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day olds prefer their native language. *Infant Behavior and Development*, *16*, 495–500.
Murty, L., Otake, T., & Cutler, A. (2007). Perceptual tests of rhythmic similarity: I. Mora rhythm. *Language and Speech*, *50*, 77–99.
Muthusamy, Y., Barnard, E., & Cole, R. (1994). Automatic language identification: A review/tutorial. *IEEE Signal Processing Magazine*, *11*, 33–41.
Navrátil, J. (2001). Spoken language recognition: A step toward multilinguality in speech processing. *IEEE Transactions on Speech and Audio Processing*, *9*, 678–685.
Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 756–766.
Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*, 1–19.
Ohala, J. J., & Gilbert, J. B. (1979). Listeners' ability to identify languages by their prosody. In: P. Léon, & M. Rossi (Eds.), *Problèmes de prosodie* (pp. 123–131). Ottawa: Didier.
Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (pp. 1–292)Massachusetts Institute of Technology1–292.
Pike, K. (1946). *The intonation of American English*. Ann Arbor: University of Michigan Press.
Ramus, F. (2002a). Acoustic correlates of linguistic rhythm: Perspectives. *Presented at Speech Prosody, 2002*, Aix-en-Provence, France.
Ramus, F. (2002b). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*.
Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of Acoustical Society of America*, *105*, 512–521.
Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265–292.
Richardson, J. A. C. (1973). *The identification by voice of speakers belonging to two ethnic groups*. Ohio State University [dissertation].
Szakay, A. (2008). *Ethnic dialect identification in New Zealand: The role of prosodic cues*. Germany: VDM Verlag.
Wagner, P., & Dellwo, V. (2004). Introducing YARD (yet another rhythm determination) and re-introducing isochrony to rhythm research. *Speech Prosody*, 227–230 [in Nara, Japan].
White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*, 501–522.
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. (2010). How stable are acoustic metrics of contrastive speech rhythm?. *Journal of Acoustical Society of America*, *127*, 1559–1569.
Willems, N. (1982). *English intonation from a Dutch point of view*. Dordrecht: Foris.