Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Automatic Detection of Brazil's Prosodic Tone Unit

David O. Johnson[1], *Okim Kang*[1]

[1] Northern Arizona University, Flagstaff, AZ, USA
David.Johnson@nau.edu, Okim.Kang@nau.edu

Abstract

This research is focused on the automatic detection of one of the fundamental elements of Brazil's prosody model, the tone unit. We compared the performance of using silent pause duration alone to delimit tone units and using relative pitch resets and slow pace (or post-boundary lengthening) along with silent pause duration to delimit them. The corpus used for the comparison is composed of 18 highly proficient speakers giving academic lectures in six varieties of English which are representative of the inner (American and British), outer (Indian and South African), and expanding (Chinese and Spanish) concentric circles of Kachru's World Englishes. The performance was compared by computing Pearson's correlation between the numbers of tone units in a trained linguist's transcription of the corpus and the numbers automatically detected by the computer. The computer detected the tone units from phone sequences identified in the audio files by a large vocabulary spontaneous speech recognition (LVCSR) program. We found including relative pitch resets and slow pace along with silent pause duration in the computer algorithm improved the correlation between the numbers of tone units in the linguist's transcription of the corpus and the numbers automatically detected by the computer from 0.935 to 0.959.

Index Terms: Brazil's prosody model, automatic speech recognition (ASR), tone unit, World Englishes, large vocabulary spontaneous speech recognition (LVCSR)

1. Introduction

This study examines automatic detection of the *tone unit* which is the fundamental element of Brazil's model of prosody [1]. Brazil's tone unit is an intonational unit. An utterance may contain one or more tone units. Even though there is no exact equivalent of Brazil's tone unit and the intonational phrase, intermediate phrase, or accentual phrase of the ToBI (tones and break indices) model, the unit of an utterance between break indices 3 and 4, which Beckman and Ayers referred to as an intonationally labelled prosodic group, is probably the most similar [45]. Brazil describes a tone unit as a segment of a discourse that a hearer can perceive as possessing a pattern of falling and rising tones that is not the same as those of tone units exhibiting other patterns of intonation. A tone unit includes one or more *prominent syllables*. A syllable is considered prominent if it has increased pitch (Hz), duration (seconds), or intensity (dB) from that of non-prominent syllables [2]. Brazil insists that prominence is associated with the syllable as opposed to the word and is different from lexical stress. Prominence is the use of heightened pitch, duration, or intensity on a syllable to highlight a word's intention or significance. The *tone choice* of the last prominent syllable and the *relative pitch* of the first and last prominent syllables define the intonation pattern of a tone unit. Brazil specified five tone choices based on the pitch contour of the prominent syllable: falling, rising, rising-falling, falling-rising, and neutral. He postulated three even gradations of relative pitch: low, mid, and high.

In prosodic analyses, applied linguists sometimes use silent pauses greater than 100 ms [3-5], 200 ms [6], or 250 ms [7] to delimit tone units. Although this technique is practical, it does not necessarily follow Brazil's definition of a tone unit. In fact a paired two sample for means t-test of the World Englishes corpus (see below) showed there was a significant difference ($t(17)=5.48$, $p<0.01$) between the number of tone units transcribed by a trained human linguist (M=55.3, σ=146.0) and the number of tone units based on 100 ms pauses (M=61.6, σ=242.0). Thus, it would seem that a more sophisticated method of detecting tone units is in order.

The purpose of this study was to explore methods of delimiting tone units that relied on more than just silent pauses longer than a specified length. In addition to pause length, two other phenomena that have been linked to prosodic boundaries are relative pitch reset and slow pace [8]. A relative pitch reset is observed as a break in the pitch. Slow pace (a.k.a., post-boundary lengthening) is the elongation of the start of a prosodic unit. In this study, we examine whether there was improvement in the correlation between the number of tone units identified by a trained linguist and the number identified by a computer program using relative pitch resets and slow pace, along with silent pauses, to delimit tone units.

We begin the paper by reviewing related research. Then, we describe the corpora we utilized, the generation of the phone sequences from the audio files in the corpus, the automatic tone unit detection algorithm, and the experimental design. Next, we report the correlation between the number of tone units identified by a trained linguist and the number identified by a computer program employing relative pitch resets and slow pace, along with silent pauses, to delimit tone units, followed by a discussion of the implications of the results. We conclude with a summary and some future research possibilities.

2. Related Research

Wagner and Watson reviewed a number of events, besides silent pauses, that have been found to signal the boundary between two intonational units [8]. Pre-boundary lengthening [9, 11-19] and post-boundary lengthening [9, 14, 20-23] are two dependable indicators of an intonational unit boundary. As well as post-boundary lengthening, new artifacts are sometimes introduced to intensify the start of a new intonational unit. Dilley et al. [24] and Redi and Shattuck-Hufnagel [25] demonstrated that speakers often employed glottal stop insertion to signal the beginning of a new

 doi: 10.21437/SpeechProsody.2016-59

intonational unit. A disjointed break in pitch, or a relative pitch reset, has also been found to mark the boundary between intonational units [26-27]. Lehiste noticed that a creaky voice is a frequent indication of the conclusion of an intonational unit in English and various other languages [9]. Intensity is a third signal for intonational boundaries in addition to pitch and duration. Chavarria et al. reported that some speakers lowered the intensity of the phones before an intonational unit boundary, but that this boundary signal was inconsistent among speakers [28]. Recently, Staples [29] manually identified Brazil's [1] tone unit boundaries using Praat [30], a computerized speech analysis program, to detect silent pauses and pitch resets.

González-Ferreras et al. [31] utilized a number of potential boundary markers to automatically detect the boundaries between intonational units including: within-word pitch range, interval from average to maximum within-word pitch, interval from minimum to average within-word pitch, change between within-word pitch and utterance pitch means, within-word intensity range, maximum and average within-word intensity range, minimum and average within-word intensity range, maximum vowel nucleus duration normalized across all vowel types in the utterance, silent pause duration, parts-of-speech tags, plus features derived from the TILT pitch contour model [32], and a pitch contour model based on Bézier parameters [33].

Chapters 10.3 and 11.1 of *The Oxford handbook of laboratory phonology* include additional information on prosodic structure [10].

Overall, it is still unclear how intonational units can be effectively made particularly in the process of automatic detection. Identification of tone units needs to be further validated empirically as well as methodologically. The current study compares the method of using silent pause to that of relative pitch resets and boundary lengthening and attempts to determine its reliability.

3. Methods

3.1. World Englishes Corpus

The World Englishes corpus is composed of speech files and orthographic transcriptions from studies on the intelligibility of different varieties of English [34-36]. The speech samples are academic lectures lasting from four to five minutes spoken by 18 English speakers. There are one female and two male speakers from each of six distinct classifications of World Englishes: the inner circle, the outer circle, and the expanding circle. The first languages of the speakers typify each of three concentric circles of World Englishes [37]. The accents of the inner circle speakers were American (California) and British (Southern England). The outer circle speakers had Indian (Hindi) and South African (Afrikaans) accents. The expanding circle of speakers spoke with Chinese (Mandarin) and Spanish (Sonora, Mexico) accents. Chinese and Spanish were chosen because they exemplified unrelated language groups [38]. Although very proficient in English, the South African, Indian, Chinese, and Mexican speakers still had a noticeable accent confirmed by eight expert raters, who provided a scalar rating of their sample lecture recordings using a five-point scale with 'easy to understand' and 'difficult to understand' as endpoints. Using mean rating scores, speakers were selected from each of the outer circle and expanding circle varieties to represent low, mid and high degrees of comprehensibility. The 18 English speakers' accent and comprehensibility were further confirmed by 48 novice college listeners.

Speakers read TOEFL CBT listening passages provided by the Educational Testing Service. The form of the TOEFL type materials in the first task was controlled for style (a professor's monologic speech), length (passages of 500-800 words), number of questions (six), and content (topics deemed appropriate for university level students). Based on statistics of item difficulty and passage familiarity, passages and items were evenly and systematically distributed across speakers of six countries, according to topics, item difficulty for each testlet, and other task features.

Each speaker was chosen for four characteristics suggested by Major et al. [38]: each sounded like a genuine speaker of the specified accent; each could effortlessly handle the lecture's terminology; each read the lecture as though they were a professional in the subject matter covered; and each had the intonation of a seasoned academic lecturer.

We chose the World Englishes corpus for this research because it represented all three concentric circles of Kachru's World Englishes, not just standard American English [37].

A trained linguist labeled the tone units by listening to the speech files and by utilizing the Multi-Speech and CSL Software to measure the length of the silent pauses [41]. A second trained linguist labeled about ten percent of the files to validate the dependability of the tone unit annotation. The two linguists reviewed any discrepancies and recommenced labeling the files until they reached agreement on the labeling. After reaching agreement, the first linguist finished labeling the remainder of the speech files alone. This technique of labeling has been extensively followed as a trustworthy method of annotation in other applied linguistics studies [4, 5, 44].

3.2. TIMIT Corpus

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) consists of ten sentences read by each of 630 speakers from eight main dialect areas of the United States [39]: New England, Northern, North Midland, South Midland, Southern, New York City, Western, and Army Brat (moved around). The sentences include two dialect sentences, 1,890 phonetically-diverse sentences, and 450 phonetically-compact sentences. The dialect sentences were devised to classify the dialect of the speakers. The phonetically-diverse sentences are responsible for a large diversity of allophonic contexts in the corpus. The phonetically-compact sentences were designed to incorporate the most common phone pairs into the corpus along with other phone pairs that the designers thought were challenging. The two dialect sentences were recited by all speakers. Each phonetically-diverse sentence was read by a single speaker. Every one of the phonetically-compact sentences was uttered by seven dissimilar speakers.

The 6,300 sentences are split into a suggested training suite of 4,620 sentences and a test suite of 1,680 sentences. The speakers in the training suite are not the same as the ones in the test suite. Each suite contains one or more female readers and one or more male readers speaking every one of the dialects. There is very little sentence duplication among the two suites. The test suite contains all the phonemes in one or more unrelated circumstances.

3.3. Phone Sequence Generation

A KALDI-based large vocabulary spontaneous speech recognition (LVCSR) program was utilized to identify the phones in the audio speech files [40]. The LVCSR was trained with the 4,620 sentences in the TIMIT training suite. The phonetically compact and diverse utterances are the reason we selected TIMIT to train the LVCSR. The trained LVSCR is capable of identifying the phones of the 1,680 sentences in the TIMIT test set with a phone error rate (PER) of only 16%.

The LVCSR signified a silent pause in the speech files with the *sil* phone. To enhance silent pause detection, the phones from the LVCSR were modified like this: 1) exchange a *f*, *k*, *t*, *n*, or *epi* phones appearing ahead of a *sil* phone less than 100 ms in extent with a *sil*, 2) replace a single consonant phone flanked by two *sil* phones more than 100 ms in length with a *sil*, 3) merge contiguous *sil* phones into one long *sil*, and 4) substitute *sil* phones exhibiting high intensity or a pitch contour with a pseudo phone, *?*, signifying an unidentified non-*sil* phone. These modifications increased the correlation between the silent pauses identified using the Multi-Speech and Computerized Speech Laboratory (CSL) Software [41] and the KALDI detected ones and from 0.508 to 0.935.

3.4. Automatic Tone Unit Detection Algorithm

The computer algorithm automatically partitioned the phone sequences into tone units by examining the silent pauses (*sil*). A tone unit was bounded by silent pauses which were lengthier than 200 ms or with a length between 150 ms and 200 ms and either a relative pitch reset or a slow pace afterwards. The upper bound of 200 ms and the lower bound of 150 ms were determined by analyzing the manual transcriptions of a typical male (Mexican) and female (South African) in the World Englishes corpus. The computer algorithm considered a relative pitch reset to have occurred, if the syllable before a *sil* had a high relative pitch and the syllable after it had a low relative pitch or vice-versa. It deemed a slow pace to have transpired, if the interval of the syllable after the *sil* was greater than the average interval of the syllable. The average interval of a syllable was computed by adding together the average interval of the phones in the syllable. The average interval of each phone was calculated separately for each utterance. The computer algorithm regarded a syllable as three phones. The value of three was worked out as follows: the average number of syllables per word is 1.5 [42]; the average number of TIMIT phones per word is 3.9 [39]; rounding 3.9/1.5 yields an average of three phones per syllable.

3.5. Experimental Design

The experiment in this study was conducted as follows. First, the phone sequences for each of the 18 speakers in the World Englishes corpus were generated as described above. Next, the tone units for the 18 speakers were detected using the simple algorithm that a silent pause longer than 100 ms delimited a tone unit. Then, the tone units were identified utilizing silent pauses, relative pitch resets, and slow pace. Finally, the numbers of tone units detected per speaker employing both methods were compared using Pearson's correlation.

4. Results

Table 1 shows the number of tone units identified by the trained linguist for each speaker and the number automatically detected utilizing both algorithms.

Table 1. *Number of tone units detected.*

Speaker	World English	Transcript	*sil* > 100 ms	Silent pauses, relative pitch resets, & slow pace
F01	SA	58	74	60
F02	A	42	55	45
F03	C	66	75	59
F04	I	47	64	47
F05	S	50	66	46
F06	B	55	59	55
M01	B	57	76	64
M02	A	35	43	36
M03	C	57	71	61
M04	A	62	78	63
M05	S	50	53	49
M06	B	54	57	53
M07	S	74	101	76
M08	I	62	72	61
M09	SA	62	67	59
M10	I	37	51	35
M11	SA	45	55	45
M12	C	82	93	76

The first column of Table 1 gives the gender (F/M) and numerical designation of the speaker. Column two provides the World English of the speaker: American (A), British (B), Indian (I), South African (SA), Chinese (C), or Spanish (S). The number of tone units from the manual transcription of the corpus is shown in column three. The fourth column gives the number of tone units identified utilizing the simple algorithm that a silent pause longer than 100 ms delineated a tone unit. The number of tone units detected employing silent pause, relative pitch reset, and slow pace is presented in the last column. The Pearson's correlation (r) between the transcript numbers of tone units and the numbers automatically detected with the simple algorithm numbers is 0.935 and the correlation between the transcript numbers and the numbers detected automatically with the sophisticated algorithm is 0.959.

5. Discussion

In this study, we first measured Pearson's correlation between the numbers of tone units labeled by a trained linguist for each of 18 speakers of six varieties of World Englishes and the numbers recognized by a computer program that delimit tone units by silent pauses greater than 100 ms. Then, we calculated the correlation between the count of tone units annotated by a trained linguist for the same 18 speakers and the count detected by a computer program that delineated tone units utilizing an algorithm that considered silent pauses, relative pitch resets, and slow pace. We found the former correlation to be 0.935 and the latter to be 0.959. This shows that examining relative pitch reset and slow pace, along with silent pauses,

can improve automatic detection of Brazil's prosodic tone unit [1], at least with the speakers in the World Englishes corpus.

In other research, we performed an inter-rater reliability study of the human and computer tone unit annotations of utterances from the Boston University Radio News Corpus (BURNC). The BURNC is a corpus of over seven hours of speech recorded by three female and four male professional radio announcers [43]. Each story spoken by a newscaster is partitioned into paragraphs of a number of sentences. We included 144 paragraphs in our inter-rater reliability study which consisted of an equal number of males and females (3) and an equal number of clean paragraphs (vs. noisy) for each speaker (24). As part of the inter-rater reliability study we compared the number of tone units identified by three trained linguists and by a computer using the algorithm described herein. We found the Pearson's correlation between each pair of linguists was 0.911, 0.841, and 0.881 and the correlation between each linguist and the computer was 0.873, 0.855, and 0.781. Although the correlations between the computer and linguists are lower than those for the World Englishes corpus, a two-tailed two-sample t-test assuming unequal variances showed the differences between the inter-linguist correlations and linguist-computer correlations were insignificant (inter-linguist correlations: M=0.88, SD=0.03; linguist-computer correlations: M=0.84, SD=0.05; $t(4)=2.78$, $p = 0.301$). This is further evidence that the computer algorithm we proposed here can automatically identify tone units as well as a trained linguist can.

The lower correlations associated with the BURNC may be due to the fact that the algorithm was tuned to the World Englishes corpus and that even though they are both corpora of non-spontaneous read text, the two corpora are significantly different. The BURNC only contains six speakers vs. 18 speakers in the World Englishes corpus; and the BURNC speakers speak only one of the six World Englishes (i.e., American) represented in the World Englishes corpus.

6. Conclusions

In this study, we found for the speakers in the World Englishes corpus that analyzing relative pitch reset and slow pace, in addition to silent pause duration, can improve, in terms of human-computer correlation, the automatic detection of Brazil's prosodic tone unit in contrast to analyzing silent pause duration alone [1].

This is important because the tone unit is the foundation of Brazil's intonational model [1]. Consequently, if the tone units are not marked correctly, none of the other elements of the model will be right either.

In this study, we investigated only three (i.e., post-boundary lengthening, relative pitch reset, and silent pause duration) of many phenomena that have been found to signal the boundaries between intonational units. An area for further study would be to explore utilizing other boundary markers, such as: the down-step phenomena which are typical within a phrase, pre-boundary lengthening, glottal stop insertion, creaky voice, lowered intensity, within-word pitch ranges (minimum to maximum, minimum to mean, and mean to maximum), variation in means of within-word pitch and utterance pitch, within-word intensity ranges (minimum to maximum, minimum to mean, and mean to maximum), vowel nucleus duration, plus features derived from the TILT pitch contour model [32], and the Escudero-Mancebo and Cardeñoso-Payo's pitch contour model built on Bézier parameters [33].

The algorithm we employed was a fixed-path algorithm which was tuned with the World Englishes corpus. Another area for further study would be to use machine learning techniques to analyze the various boundary markers as input features to demarcate tone units. Machine learning is a branch of artificial intelligence that trains computer programs with known inputs (e.g. boundary markers) and outputs (tone unit boundaries) to recognize patterns so that they can recognize the same patterns in unknown data.

The results reported in this paper reaffirm that utilizing additional boundary markers along with silent pause duration can improve the automatic detection of Brazil's tone units over using silent pause duration alone [1].

7. References

[1] D. Brazil, *The communicative value of intonation in English*, Cambridge, England: Cambridge University Press, 1997.

[2] D. M. Chun, *Discourse intonation in L2: From theory and research to practice (Vol. 1)*, John Benjamins Publishing, 2002.

[3] J. Anderson-Hsieh and H. Venkatagiri, "Syllable duration and pausing in the speech of intermediate and high proficiency Chinese ESL speakers," *TESOL Quarterly*, 28, 807–12, 1994.

[4] O. Kang, D. Rubin, and L. Pickering, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English," *The Modern Language Journal*, 94(4), 554-566, 2010.

[5] O. Kang, "Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency," in *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference, Aug. 2012, 10-15, Ames, IA: Iowa State University*, 2013.

[6] J. T. Zeches and K. M. Yorkston, "Pause structure in narratives of neurologically impaired and control subjects," *Clinical Aphasiology*, 23, 155–4, 1995.

[7] R. Towell, R. Hawkins, and N. Bazergui, "The development of fluency in advanced learners of French," *Applied Linguistics*, 17, 84–119, 1996.

[8] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, 25(7-9), 905-945, 2010.

[9] I. Lehiste, "Phonetic disambiguation of syntactic ambiguity," *The Journal of the Acoustical Society of America*, 53(1), 380-380, 1973.

[10] A. C. Cohn, C. Fougeron, and M. K. Huffman, eds. The Oxford handbook of laboratory phonology. OUP Oxford, 2011.

[11] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *the Journal of the Acoustical Society of America*, 90(6), 2956-2970, 1991.

[12] S. Shattuck-Hufnagel and A. E. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of Psycholinguistic Research*, 25(2), 193-247, 1996.

[13] J. Edwards, M. E. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *the Journal of the Acoustical Society of America*, 89(1), 369-382, 1991.

[14] C. Fougeron, and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," *The journal of the acoustical society of America*, 101(6), 3728-3740, 1997.

[15] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple phrasal boundaries," *Journal of Phonetics*, 26, 173-199, 1998.

[16] D. Byrd and E. Saltzman, "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," *Journal of Phonetics*, 31(2), 149-180, 2003.

[17] R. Berkovits, "Durational effects in final lengthening, gapping, and contrastive stress," *Language and Speech*, 37(3), 237-250, 1994.

[18] A. E. Turk and L. White, "Structural influences on accentual lengthening in English," *Journal of Phonetics*, 27(2), 171-206, 1999.

[19] D. Byrd, J. Krivokapić, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects," *The Journal of the Acoustical Society of America*, 120(3), 1589-1599, 2006.

[20] S. A. Jun, *The phonetics and phonology of Korean prosody*, Doctoral dissertation, The Ohio State University, 1993.

[21] L. M. Lavoie, "Consonant strength: Phonological patterns and phonetic manifestations," *Psychology Press*, 2001.

[22] T. Cho, "The effects of prosody on articulation in English," *Psychology Press*, 2002.

[23] P. Keating, T. Cho, C. Fougeron, and C. Hsu, "Domain-initial strengthening in four languages," *Phonetic interpretation: Papers in laboratory phonology VI*, 143-161, 2004.

[24] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, 24(4), 423-444, 1996.

[25] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, 29(4), 407-429, 2001.

[26] J. R. de Pijper and A. A. Sanderman, "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues," *The Journal of the Acoustical Society of America*, 96(4), 2037-2047, 1994.

[27] H. Truckenbrodt, "Upstep and embedded register levels," *Phonology*, 19(01), 77-120, 2002.

[28] S. Chavarria, T. J. Yoon, J. Cole, and M. Hasegawa-Johnson, "Acoustic differentiation of ip and IP boundary levels: Comparison of L-and LL% in the Switchboard corpus," in *Speech Prosody 2004, International Conference*, 2004.

[29] S. Staples, *Linguistic characteristics of international and U.S. nurse discourse*, Doctoral dissertation, Northern Arizona University, 2014.

[30] P. Boersma and D. Weenink, *Praat: doing phonetics by computer (version 5.3.83)*, [Computer program]. Retrieved August 19, 2014.

[31] C. González-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual and V. Cardeñoso-Payo, "Improving automatic classification of prosodic events by pairwise coupling," *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(7), 2045-2058, 2012.

[32] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the acoustical society of America*, 107(3), 1697-1714, 2000.

[33] D. Escudero-Mancebo and V. Cardeñoso-Payo, "Applying data mining techniques to corpus based prosodic modeling," *Speech Communication*, 49(3), 213-229, 2007.

[34] O. Kang, "ESL learners' attitudes toward pronunciation instruction and varieties of English," in *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference. Ames, IA: Iowa State University*, 105-118, 2010.

[35] O. Kang, M. Moran, and R. Thomson, "Pronunciation features of intelligible speech among different varieties of world Englishes", presentation *at Pronunciation and Second Language Learning and Teaching Conference, Santa Barbara, CA, September 5-6, 2014*.

[36] O. Kang, R. Thomson, and M. Moran, "Intelligibility of Different Varieties of English: The Effects of Incorporating "Accented" English into High Stakes Assessment," presentation at *American Association of Applied Linguistics Conference, Toronto, ON, Canada, March 21–24, 2015*.

[37] B. B. Kachru, *The other tongue: English across cultures*, University of Illinois Press, 1992.

[38] R. C. Major, S. F. Fitzmaurice, F. Bunta, and C. Balasubramanian, "The effects of nonnative accents on listening comprehension: Implications for ESL assessment," *TESOL quarterly*, 36(2), 173-190, 2002.

[39] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, 93, 27403, 1993.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, ... and K. Vesely, "The Kaldi speech recognition toolkit," in *Proceedings of Automatic Speech Recognition and Understanding Workshop*, 1-4, 2011.

[41] KayPENTAX, *Multi-Speech and CSL Software*, Lincoln Park, NJ: KayPENTAX, 2008.

[42] S. Sakti, K. Markov, S. Nakamura, and W. Minker, *Incorporating Knowledge Sources into Statistical Speech Recognition*, (42), Springer, 2009.

[43] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, *The Boston University radio news corpus*, Linguistic Data Consortium, 1-19, 1995.

[44] L. Pickering. *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants*, Doctoral dissertation, University of Florida, 1999.

[45] M. E. Beckman and G. Ayers. *Guidelines for ToBI labelling*. The OSU Research Foundation, 3, 1997.