AUTOMATIC SYLLABLE STRESS DETECTION USING PROSODIC FEATURES FOR PRONUNCIATION EVALUATION OF LANGUAGE LEARNERS

Joseph Tepperman and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
Integrated Media Systems Center
Department of Electrical Engineering-Systems
Viterbi School of Engineering
University of Southern California
http://sail.usc.edu/
tepperma@usc.edu, shri@sipi.usc.edu

ABSTRACT

A robust language learning system, designed to help students practice a foreign language along with a machine tutor, must provide meaningful feedback to users by isolating and localizing their pronunciation errors. This paper presents a new technique for automatic syllable stress detection that is tailored for language-learning purposes. Our method, which uses basic prosodic features and others related to the fundamental frequency slope and RMS energy range, is at least as accurate as an expert human listener, but requires no human supervision other than a pre-defined dictionary of expected lexical stress patterns for all words in the system's vocabulary. Optimal feature choices exhibited an 87-89% accuracy compared with human-tagged stress labels, exceeding the inter-human agreement commonly held to be about 80%.

1. INTRODUCTION

Awareness of proper lexical stress is very important to students of a foreign language. In English, for instance, misplaced syllabic stress can alter a word's part of speech (in the case of "rebel" or "insult") or even change the word's meaning entirely (as with "content" or "contract"). So any interactive computer program for language learners needs to be able to automatically detect a non-canonical stress pattern at least as well as a human tutor would.

In the past, prosodic features have been used to successfully identify syllabic stress, but usually only with some catch that renders such methods unsuitable for most language learning applications. In [8], stress (or "prominence") is a strict two-class problem, applied only to individual syllables taken out of their word context. A language learning tool, though, will be interested in classifying only one syllable per word as the location of primary stress, rather than classifying each syllable individually as stressed/unstressed. The language learning tool implemented in [3] can detect misstressed syllables only by comparing the student's pronunciation with that of a master signal, pre-recorded and hand labeled for prosodic information. This, however, requires considerable human input (and hence, scalability is an issue), and also limits the program's vocabulary to those words that have previously been recorded and tagged.

The method outlined in this paper requires no human supervision in terms of marking speech data with stress information. Instead, it uses a dictionary of canonical word transcriptions for forced phonetic alignment and subsequent feature extraction and classification. Since this is designed for language learning modules, in which registered users' pronunciations are evaluated based on utterances spoken after machine prompts, it's safe to assume that the aligner has prior knowledge of said prompts (and their expected transcriptions), as well as perhaps some modicum of background meta-information on the speaker with which to optimize pronunciation evaluation.

The data we used for these experiments came from the ISLE Corpus compiled at the University of Leeds [1]. These recordings consist of 46 adult Intermediate British English learners who are native speakers of either Italian or German – 23 speakers of each. Utterance prompts were complete sentences written by design to highlight certain difficulties English learners typically encounter, both in phonemic pronunciation and in recognizing variations in primary lexical stress (e.g. "project" when used as a noun vs. when used as a verb). The recordings were automatically tagged for canonical forms by a forced-aligner, then corrected to reflect the speaker's pronunciation by a team of five linguists, who also added labels for each word's syllable of primary acoustic stress (compared to the canonical).

2. PROSODIC FEATURES

2.1 The Syllable Nucleus

The syllable as a lexical unit is well defined, orthographically speaking. Open any standard dictionary and there you'll find every word parsed by fixed linguistic rules into its component syllables. But as an acoustic unit of speech, the exact phonemic boundaries of every syllable vary depending on the rate of speaking and rhythmic flow of pronunciation. Because syllabic durations are difficult to obtain from the speech signal itself (as explained in [8]), we might generate a dictionary using the tsylb2 automatic syllable parser [7], which takes an arbitrary phonetic transcription and returns its most likely syllabic concatenation, but with a list of two or three close but differing results corresponding to likely variations in speaking rate. The word "syllable," for instance, might be pronounced 's_ih' 'l_ax' 'b_ah_l' at normal speaking rates, but as 's_ih_l' 'ax_b' 'ah_l' by an abnormally fast speaker.

For these reasons, we have followed [8,9] and opted not to extract prosodic features from dubious syllable units, but rather from the syllable nucleus, the essential vowel center of a syllable (the boundaries of which can easily be obtained from forced alignments). Results stated in [8,9] indicate that syllabic stress is highly correlated with the prosodic features derived from syllable nuclei.

The decision to use syllable nuclei in place of the whole syllable itself makes sense in light of the prosodic features traditionally used to detect stress. In choosing the syllable nucleus as the area of interest, we are exploiting three basic components of prosody – fundamental frequency (f0), energy, and duration. Vowels tend to be the most telling piece of overall syllable characteristics, because the surrounding consonants are typically shorter, quieter, and less reflective of subtle spectral transitions that indicate the presence of syllabic stress. In a word like "media," the /ax/ vowel at the end is an entire syllable unto itself. So, really now, for stress detection the syllable and its nucleus are interchangeable as areas of acoustic interest.

2.2 Choice of Features

Just what is syllable stress? Other papers on this topic ([3,7,8]) confound the terminology by using "prominence," "stress," and "accent" sometimes interchangeably, but often in referring to similar and easily confusable phenomena. In [8], an accented event is defined as one that "exhibits a rise followed by a fall profile," presumably in pitch. Whereas a stressed syllable is linked only to an increase in duration and energy, but not pitch. Prominence, then, is an all-purpose term encompassing either a pitch-accented or stressed syllable.

So the "stressed" term used throughout this paper, referring to a syllable "perceived as standing out from its environment" in the form of primary lexical emphasis, is closest to [8]'s definition of "prominence," since we define it to encompass prosodic features related to pitch, energy, and duration. According to [7], "it has been proven that f0 is not a reliable correlate of stress." True, it is not nearly as reliable as energy or duration (at least not in [7]'s study of Dutch and American English). For instance, in English, a rise in pitch at the end of a question does not necessarily correspond with syllable or word-level stress. But for students learning English as a foreign language, the inclusion of pitch-related features may prove indispensable to detecting their syllable stress placement, especially if their native tongue is one like Mandarin, in which pitch dominates as the feature that dictates both word meaning and syllabic stress. And as far as choosing one of several syllables as the location of primary stress within a word (as our current investigation aims to do), features related to the fundamental frequency - though perhaps not the simple f0 values - do significantly improve detection accuracy.

The three basic prosodic features chosen for baseline experiments were mean values of f0 and energy over the nucleus (normalized by the respective mean values over the entire word) and the nucleus duration (normalized by the mean nucleus duration over all syllables in that word). Normalizing in this way preserves the word context information and is in keeping with the fundamental idea that we're not so much interested in classifying each syllable separately, but rather we intend to compare characteristics of all syllables in a word and choose one as the location of primary stress.

Though we are using the relatively unreliable f0 value as a feature, we also included several features related to the f0 slope, which proved to be closely correlated with syllabic stress. The importance of these slope-derived features is in their potential to capture higher-level pitch information, to model the rapid rate of f0 and energy changes that correlate with stress but are in some sense independent of the mean f0 value. These features are inspired by the ones used for pronunciation evaluation and speaker recognition in [9] and [6], respectively. They are:

- the mean f0 slope over the nucleus, divided (which is to say normalized) by the mean slope over the entire word
- the total number of rising and falling frames in the nucleus, normalized by the total number of frames in the word
- the number of intra-nucleus changes from rising to falling slope (or vice versa) between adjacent frames, normalized by the total frames in the word
- the pitch pseudo-slope (the last value minus the first value) over the nucleus, divided by the pseudo-slope of the whole word

For reasons similar to those of the slope-derived features, we also included the following range features:

- f0 range over the nucleus, divided by the f0 range over the whole word
- energy range over the nucleus, divided by energy range over the word

Including these slope- and range-related features serves to make our model of syllabic "stress" a combination of [8]'s close distinctions between stress and pitch-accent, which we argue is necessary for language learning purposes.

2.3 Feature Extraction and Processing

As stated above, the syllable nuclei durations were derived from automatic results of forced alignments based on transcriptions of each recording's utterance prompt. The phonemic boundaries (in milliseconds) the ISLE corpus has presented as text files grouped into sentence utterances, labeled with expected and transcribed pronunciations and stress patterns. So it was possible for us to incorporate higher-level contextual information to optimize normalization of our syllable duration feature, which was necessary especially because we used the syllable nucleus in place of the syllable itself.

[3] presents a list of linguistic rules with which to further normalize vowel durations based on that vowel's word- and phrase-level context. These rules seem to be derived from empirically calculated average duration trends in pronunciation. The ones we used are as follows:

- Divide by 2 all vowel durations in prepausal words
- Divide by 1.5 all vowel durations in content words
- Divide by 2 all vowel durations preceding voiced fricatives
- Divide by 1.5 all word-final vowel durations
- Divide by 1.25 all vowel durations preceding a voiced stop
- Divide by 0.5 all vowel durations preceding an unvoiced stop

The above rules apply only for the nuclei of syllables expected to hold primary stress, which we may assume the classifier knows from a dictionary. There is one more rule, though, for the syllables expected to be canonically unstressed:

- Divide by 0.5 all vowels expected to be unstressed

For purposes of comparison, we also kept the original durations before the contextual rules were applied.

The f0 and RMS energy values were obtained using the ESPS get_f0 pitch tracking method with the default frame length of 10 msec and pitch range from 60 to 400 Hz.

As recommended in [10], for all f0-related features we ignored the last four frames of any word-final prepausal vowel, so as to avoid the "boundary tones" that denote the end of a phrase. The mean slope and pseudo-slope features were taken as absolute values, because the sign might have flipped after normalization. And we used log-values of all features, so as to better fit them with Gaussian distributions.

3. EXPERIMENTS AND RESULTS

3.1 A Two-class Problem

We began the classification process by considering this a two-class problem: though we included the contextual rules listed above in Section 2.3, we still started by classifying each syllable individually as stressed or unstressed, without regard for within-word information (i.e. without considering that every word should have one and only one syllable of primary stress). This is because to classify the primary stress of an N syllable word is really an N-class problem, for which we would have to separately classify two-syllable words, three-syllable words, etc., assuming we don't build a unique model for every possible stress pattern of every word! So, to limit classification complexity, we initially considered only two possible classes and one generalized classifier.

For training data, we used 13 native Italian speakers (a total of 7086 syllable nuclei, taken only from polysyllabic words), and 12 native German speakers (7878 syllable nuclei instances), all taken from the ISLE corpus described above in Section 1. Classifiers for the Italian and German students were trained and tested separately, since we may assume the classifier has prior knowledge of the registered student's native language. The test set was comprised of the remaining 10 Italian speakers (6667 nuclei from 2914 polysyllabic words) and the remaining 11 German speakers (7021 nuclei taken from 3065 polysyllabic words). The ratio of unstressed to stressed vowels in each of the training and test sets was about 1.29.

The models for stressed and unstressed syllable nuclei were built in MATLAB as Gaussian mixtures using the PRTools Toolbox [4]. The classifier chosen was a quadratic Bayes discriminant function.

3.2 Incorporating Word Information for Post-Classification

After individually classifying every syllable nucleus as stressed or unstressed, we sought to improve accuracy by including information about intra-word stress results. By definition, no word can have more or less than one syllable of primary stress, and the ISLE corpus is labeled accordingly. So in the words for which our two-class classifier assigned more than one primary stress, we kept the one with the best posterior probability returned by the classifier, and post-classified the other ones as unstressed. And in words with no stressed syllable results, we chose the one unstressed syllable with the worst posterior probability and post-classified it as stressed. Then, after this post-classification step, we retested our individual nuclei results, and also counted how many complete words had had all their syllables properly tagged (an inter-word accuracy to compare with the inter-syllable accuracy after the word information had been applied).

Results using different feature sets are shown in Table 1. All results listed include the seven contextual rules for normalizing vowel durations, because without these rules results were significantly less accurate (by anywhere from 5 to 10%), and incorporating the word information actually worsened inter-word accuracy.

The errors were split almost exactly evenly between missed detections and false alarms.

4. DISCUSSION

From Table 1 we can see that the classifier performed slightly better overall for the Germans than for the Italians. This might be due to the fact that the contextual rules for normalizing durations in [3] were taken from a study in American English, and German is linguistically more closely related to English than Italian is. However, testing the German speakers on the Italian-trained models (and vice-versa) did not result in a significant decline in accuracy. This seems to indicate that, with more diverse training data, one should be able to generalize these models for all English learners, regardless of their native language (or at least define models in broader linguistic groups).

As we expected, the post-classification done in incorporating word information always improved accuracy for individual syllables. But the human experts did not tag each syllable individually, without regard for other syllables in the same word. No, they listened to each word separately and picked one syllable as the location of primary stress. So in the end, the best measure of our method's performance is really the word accuracy ratings – the percentage of words in which all syllables were classified correctly. Now inter-human agreement in linguistic labeling is commonly held (by [8,9]) to be about 80%. So, by Table 1, even a few of our sub-optimal feature selections performed as well as a human labeler would. And our results using baseline features were comparable with that of similar features employed in [7,8].

The inter-word accuracy metric (Table 1's "word" columns)

	Accuracy (%)					
	Italian			German		
	syllable	w/ word info	word	syllable	w/ word info	word
3 basic features	75.63	76.51	83.39	80.26	82.00	87.80
ranges	56.74	57.34	64.96	61.96	62.46	69.69
slopes	62.82	66.91	73.10	67.43	69.15	75.60
basic + ranges	78.16	78.31	83.87	82.07	82.25	87.41
basic + slopes	82.48	84.01	87.61	85.49	85.73	89.14
all 10 features	82.57	83.17	86.75	85.57	85.81	88.81

Table 1. Syllable stress detection accuracy for individual syllables with and without word information, and inter-word accuracy based on the percentage of complete words in which all syllables were post-classified correctly.

always yielded better results than the syllable accuracy with word info metric. This must be because most of the words were only two syllables in length, and words with more syllables are more likely to have their overall stress patterns classified incorrectly (as the individual syllable classification error adds up). But, after post-classification, a given word can have no more (or less) than two syllables classified incorrectly. When a two-syllable word is incorrect, all (both) syllables will be misclassified. So shorter words, which comprise the majority of the corpus, will perform worse for inter-syllable accuracy, but better for inter-word accuracy.

As far as the features go, the baseline experiments using only mean f0, mean energy, and duration already yielded a word accuracy rate of better than 80%. Adding features related to the f0 slope and the f0 and energy ranges was necessary only to push the individual syllable accuracy above 80%, in keeping with the rating convention of [8]. Though the range-related features did add some improvement over the baseline, the classifier performed better when only the baseline and slope-related features were used. But, that's really just an improvement in agreement with a human labeler, so any increase in accuracy beyond 80% is redundant.

As a supplementary experiment, we calculated the classifier output's accuracy when compared with the canonical stress pattern (not the human-tagged stress labels), by way of generating some kind of crude overall pronunciation score for each speaker (a true score would include more than just accuracy of stress placement). We also calculated comparable objective "human" scores based on the agreement between the hand-tagged stress data and the canonical stress marks. The correlation between the automatic scores and the human scores for the 10 speakers in the Italian test set was 0.7998 for syllables, 0.8087 for words, which beats the inter-human correlation in assigning general pronunciation scores to speakers in this corpus (it was no better than 0.7, as reported in [1]). This speaks well for our method's applicability to pronunciation evaluation.

The native Italian speakers in the ISLE corpus frequently mispronounced /uh/ with rounded lips as /uw/ [1]. Similarly, the native German speakers often pronounced their /z/ unvoiced as /s/. So future work along these lines would be well-served to incorporate articulatory features into the pronunciation scoring scheme, for automatic detection of typical pronunciation errors and localization of the exact type of mistake (manner, voicing, place, rounding, etc.) made in the misarticulation. It might also be interesting to see how these features perform in detecting word-level stress within an entire utterance, though for language-learning purposes syllable-level stress is far more crucial to effective communication.

5. ACKNOWLEDGEMENTS

This project was supported in part by the DARWARS initiative sponsored by the US Defense Advanced Research Projects Agency and a grant from the National Science Foundation.

6. REFERENCES

[1] E. Atwell, P. Howarth, C. Souter, "The ISLE Corpus: Italian and German Spoken Learner's English," *ICAME Journal*, Vol. 27, pp. 5-18, 2003.

[2] R. Bates and M. Ostendorf, "Modeling Pronunciation Variation in Conversational Speech Using Prosody," ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language, September 14-15, 2002.

[3] R. Delmonte, M. Petrea, and C. Bacalu, "SLIM Prosodic Module for Learning Activities in a Foreign Language," *Proc. ESCA*, Eurospeech97, Rhodes, Vol. 2, pp. 669-672.

[4] Duin, R.P.W., *PRTools 3.1.7*, Delft University of Technology, the Netherlands, http://www.prtools.com/, 2002.

[5] Fisher, Bill, *tsylb2-1.1 syllabification software*, National Institute of Standards and Technology, http://www.nist.gov/speech/, 1996.

[6] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, B. Xiang, "Using Prosodic and Conversational Features for High-Performance Speaker Recognition," *Johns Hopkins University Workshop 2002*.

[7] A. Sluijter and V. van Heuven, "Acoustic Correlates of Linguistic Stress and Accent in Dutch and American English," *Proc. ICSLP'96*, Philadelphia, pp. 630-633.

[8] F. Tamburini, "Prosodic Prominence Detection in Speech," *Proc. ISSPA2003*, Paris, pp. 385-388.

[9] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, K. Sonmez, "Evaluation of Speaker's Degree of Nativeness Using Text-Independent Prosodic Features," *Proc. Of the Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[10] C. Wightman and M. Ostendorf, "Automatic Labeling of Prosodic Patterns," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, pp. 469-481, 1994.