Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 69 (2015) 31–45

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems

Luciana Ferrer *, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, Kristin Precoda

Speech Technology and Research Laboratory, SRI International, CA, USA

Received 15 May 2014; received in revised form 23 January 2015; accepted 5 February 2015
Available online 18 February 2015

Abstract

We present a system for detection of lexical stress in English words spoken by English learners. This system was designed to be part of the EduSpeak® computer-assisted language learning (CALL) software. The system uses both prosodic and spectral features to detect the level of stress (unstressed, primary or secondary) for each syllable in a word. Features are computed on the vowels and include normalized energy, pitch, spectral tilt, and duration measurements, as well as log-posterior probabilities obtained from the frame-level mel-frequency cepstral coefficients (MFCCs). Gaussian mixture models (GMMs) are used to represent the distribution of these features for each stress class. The system is trained on utterances by L1-English children and tested on English speech from L1-English children and L1-Japanese children with variable levels of English proficiency. Since it is trained on data from L1-English speakers, the system can be used on English utterances spoken by speakers of any L1 without retraining. Furthermore, automatically determined stress patterns are used as the intended target; therefore, hand-labeling of training data is not required. This allows us to use a large amount of data for training the system. Our algorithm results in an error rate of approximately 11% on English utterances from L1-English speakers and 20% on English utterances from L1-Japanese speakers. We show that all features, both spectral and prosodic, are necessary for achievement of optimal performance on the data from L1-English speakers; MFCC log-posterior probability features are the single best set of features, followed by duration, energy, pitch and finally, spectral tilt features. For English utterances from L1-Japanese speakers, energy, MFCC log-posterior probabilities and duration are the most important features.

Keywords: Computer-assisted language learning; Lexical stress detection; Mel frequency cepstral coefficients; Prosodic features; Gaussian mixture models

1. Introduction

Lexical stress is an important component of English pronunciation, as English makes a greater use of stress than many other languages. To understand spoken words, native[1] speakers of English rely not only on the pronunciation of sounds, but also on the stress patterns. Using the incorrect stress pattern can greatly reduce a speaker's intelligibility. This poses a big problem for English learners, especially for native speakers of languages that have more consistent lexical stress patterns or have different ways of incorporating timing and rhythm. This is especially

* Corresponding author at: Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

[1] A note on nomenclature: Throughout the paper we will use the word "native" to refer to the L1 of a speaker and, also, to whether the language being spoken is the speaker's L1. Hence, the phrase "native English speakers" refers to L1-English speakers, the phrase "native Japanese speakers" refers to L1-Japanese speakers, and the phrase "non-native English speakers" refers to speakers with L1 other than English. Furthermore, we will call "native data" any data where the language spoken is the same as the L1 of the speakers, and "non-native data" any data where the language spoken is not the same as the L1 of the speakers. When no language is specified, native and non-native refer to native English and non-native English (data or speakers), respectively.

http://dx.doi.org/10.1016/j.specom.2015.02.002
0167-6393/

true for native Japanese speakers learning English: in Japanese, the rhythm is more regular and syllables are more similar in prominence than in English. Computer-assisted language learning (CALL) software can then greatly benefit from the ability to provide feedback about stress pronunciation to the user.

A large variety of automatic systems that use different features and modeling techniques to classify stress have been proposed in the literature. Unfortunately, as we explain below, many of them are unsuitable for use in CALL systems because the assumptions they make do not apply to language learners. Many others were not tested on non-native speakers of the language for which the system was trained and, hence, their suitability for CALL systems is unknown.

Most proposed stress classification systems are based on prosodic features like pitch, energy and duration, which are normalized in different ways to make them independent of the speaker's baseline pitch, the channel volume, the speech rate and so on. Measurements are generally obtained only over the nucleus for each syllable. Examples of this kind of segmental features can be found in several papers (Tepperman and Narayanan, 2005; Chen and Wang, 2010; Deshmukh and Verma, 2009; Chen and Jang, 2012; Verma et al., 2006; Zhu et al., 2003). Spectral features, on the other hand, have been rarely used for stress detection. Li et al. (2007) and Lai et al. (2006) propose similar systems using mel-frequency cepstral coefficients (MFCCs) modeled by hidden Markov models (HMMs). Both papers address the problem of detecting English sentence-level stress rather than word-level stress and test only on data from native English speakers.

Modeling techniques for stress detection vary widely and include decision trees (Deshmukh and Verma, 2009), Gaussian mixture models (GMMs) (Tepperman and Narayanan, 2005; Chen and Jang, 2012), support vector machines (Deshmukh and Verma, 2009; Chen and Wang, 2010; Zhao et al., 2011), deep belief networks (Li et al., 2013), and HMMs (Lai et al., 2006; Li et al., 2007; Ananthakrishnan and Narayanan, 2005). In many cases, the task of stress detection is defined as the problem of locating the single primary stressed syllable in a word. Under this assumption, modeling techniques can make a single decision per word – rather than one decision per syllable – using features extracted from all syllables in the word (Chen and Wang, 2010; Chen and Jang, 2012) or obtain syllable-level scores and then choose the syllable with the largest score as the primary stress location (Tepperman and Narayanan, 2005; Zhao et al., 2011). Furthermore, some techniques require that words have correct phonetic pronunciation in order to make a stress level decision (Chen and Jang, 2012). Finally, the task of labeling each syllable in an utterance from a non-native English speaker as unstressed, primary stressed or secondary stressed is an extremely complex one. In our database, the observed disagreement for native Japanese children speaking English across three annotators is, on average, 21% (corresponding to an agreement of 79%). Given this difficulty, some researchers simplify the labeling task by asking annotators to assign "correct" versus "incorrect" labels to each word rather than actual stress pronounced on each syllable (Deshmukh and Verma, 2009; Verma et al., 2006) or by labeling only the location of the primary stress (Tepperman and Narayanan, 2005; Chen and Jang, 2012). Many of these modeling and labeling assumptions are inappropriate for language learners who will most likely mispronounce both phones and stress within a word and might pronounce more than one syllable with primary stress.

We describe a novel system for lexical stress feedback intended for use by native Japanese children learning English. We expect the learners to pronounce sounds poorly and to pronounce most syllables with more prominence than native English speakers would. In fact, according to our phonetician's annotations, in our Japanese children's database around one third of the incorrectly stressed words have primary stress in at least two syllables. Therefore, our system must allow more than one syllable with primary stress in a word. Furthermore, phonetic and stress pronunciations are tied together; pointing out a stress mistake might go a long way toward fixing the phonetic mistakes, and conversely. For this reason, we do not wish to assume correct phonetic pronunciation before giving feedback about the stress pronunciation.

The proposed system is designed to approximate the decisions a phonetician would make about the stress level pronounced for every syllable in a word. For the Japanese children data, the system is evaluated against decisions made by annotators. The goal is to approximate those decisions as well as possible. Hence, the most natural approach would be to train such a system using data from the same population of Japanese children speaking English. This way, the model would describe the stress level as pronounced by this population of speakers. Nevertheless, since the stress labeling task is costly and agreement is low, little amount of data is available with reliable labels for training the system. For this reason, we propose to use utterances from native English speakers to train our system. For this data, stress labels are obtained automatically, assuming that native English speakers pronounce stress in a predictable manner for selected words according to a dictionary. While this approach results in models that represent stress as pronounced by native English speakers, we show that it results in good performance on the Japanese children's data. Matched Japanese children's data can then be used to fine-tune the system through adaptation of the models.

The decisions made by the system are meant to be used as a tool within CALL software. The software could be designed to only correct the speaker when the stress mistake would result in intelligibility problems (for example, when the meaning of the word depends on the stress pattern). On the other hand, the software could aim at achieving native-like pronunciation, correcting the speakers every time they make a mistake, regardless of whether this would

cause intelligibility problems or not. The specific exercises assigned by the software and the way the stress decisions output by the system are used to give feedback to the user are not the subject of this work.

Our approach to stress detection models features based on duration, pitch, energy, spectral tilt and MFCC-based measurements over the syllable nuclei. The first three types of features are commonly used in the literature on stress detection, with spectral tilt and MFCC-based measurements being less common. A novel aspect of our system is the successful integration of spectral information (MFCCs and spectral tilt) and prosodic (duration, pitch and energy) information. It is reasonable to assume that, given the phonetic pronunciation mistakes made by language learners, spectral features would fail to carry robust stress information, especially when models are learned using data from native English speakers. Nevertheless, as we will see, we find significant gains from the inclusion of this information in the system for both native and non-native English data. We propose to use GMMs for stress modeling and show that, when adaptation techniques are used to obtain robust models, this method outperforms decision tree and neural network methods.

Finally, another novel aspect of our proposed system is the way it makes the final decisions. Stress detection systems for language learning commit two types of errors: false corrections, where the learner is corrected when he actually pronounced stress correctly; and missed corrections, where a mistake made by the learner goes uncorrected. We believe the first kind of error is much more bothersome to the learner than the second one. A student that is constantly corrected when he feels he has done it right will be likely to stop using the system. Hence, we want to be able to set the system to operate with a certain maximum level of false corrections, even if this implies an increase in the rate of missed corrections. To our knowledge, all papers in the area of stress detection report results on hard decisions made by the system, with most papers reporting a single accuracy number (Verma et al., 2006; Ananthakrishnan and Narayanan, 2005; Chen and Wang, 2010; Lai et al., 2006; Tepperman and Narayanan, 2005). Our system generates posterior probabilities that are used to make the final decisions with thresholds that are chosen according to the desired maximum level of false corrections. We report the percent of missed corrections at a false correction level of 5%, along with the more traditional error rates.

As a consequence of the work presented in this paper, stress classification capabilities were integrated into SRI's CALL toolkit, EduSpeak®. This makes EduSpeak one of very few commercially available CALL toolkits with stress classification capabilities.

The rest of the paper is organized as follows. Section 2 describes the system architecture, including the features, modeling technique and decision making. Section 3 describes the datasets used for the experiments; annotation statistics; performance metrics; and, finally, detailed results on the proposed system, including results from feature selection experiments. Finally, Section 4 gives our conclusions and future work.

2. System description

Our proposed stress detection system is designed to perform well on non-native English utterances, satisfying the following constraints: (1) the system should predict stress for each syllable using three levels: unstressed, primary and secondary stress; (2) the system should not assume a single syllable per word has primary stress; (3) the system should be word-independent (no previous knowledge of the words of interest can be used); (4) the system should not rely on good phonetic pronunciations; and (5) there should be a way to maintain the percent of false corrections below a certain threshold. The features and system architecture defined in the following sections satisfy these constraints.

A flowchart of the system is given in Fig. 1. The following sections describe the different steps in this figure in detail.

2.1. Features

Features are extracted over the nucleus of each syllable. Five types of segmental features are defined based on duration, pitch, energy, spectral tilt and MFCCs. All features undergo some type of normalization to make them as independent as possible of characteristics that might confound the classification of stress, like the channel, the speech rate, the baseline pitch of the speaker, and so on. We perform all normalizations at the word level. This way, syllable-level features are all relative to the mean values found in the word.

As already mentioned, we wish to design a system that works well even when the word is incorrectly pronounced. Only extreme mispronunciations, in which a full syllable was deleted, were discarded from our database during labeling. Given this type of data, we have found in preliminary experiments that vowel-dependent modeling or normalization does not lead to significant gains, even when training on matched data. For this reason, vowel-dependent modeling or normalization is not performed by our system. This result contradicts previous papers on the topic (Deshmukh and Verma, 2009; Oxman and Golshtein, 2012). We believe the likely reason for this discrepancy is that our children's database (see Section 3.1) has a very high rate of pronunciation mistakes while the Indian and Hebrew databases in those papers are likely to have better pronunciation quality.

2.1.1. Phone-level alignments

In order to locate the vowels within the waveforms, we run EduSpeak (Franco et al., 2000; Franco et al., 2010), SRI International's automatic speech recognizer (ASR) and pronunciation scoring toolkit for language learning

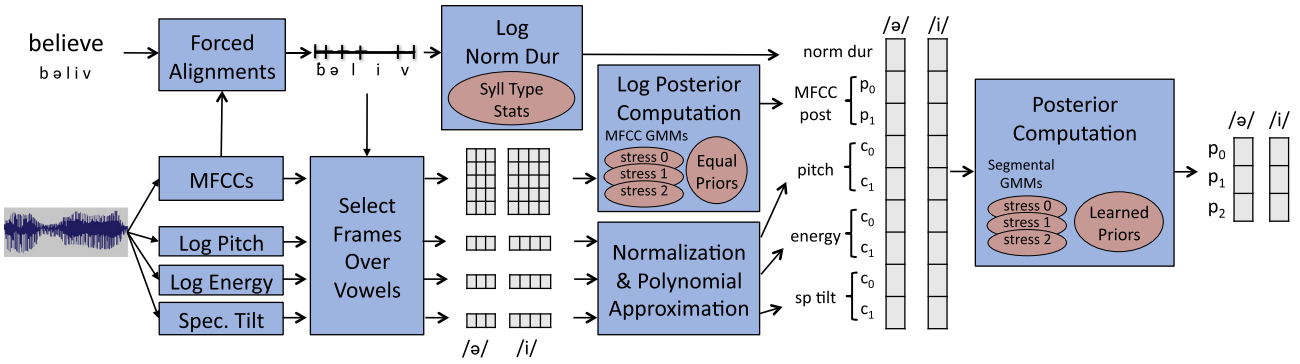

Fig. 1. Proposed stress detection system. The inputs to the system are a speech waveform corresponding to a single word, and its canonical pronunciation. Frame-level MFCCs, and frame-level pitch, energy and spectral tilt signals are estimated from the waveform. Forced alignments are created using the MFCCs and the phonetic transcription for the word. The resulting alignments are used to compute the normalized duration features and to constrain the MFCC, pitch, energy and spectral tilt frame-level features to the regions over the vowels. For each vowel, a polynomial approximation of order 1 is computed from the pitch, energy and spectral tilt normalized values, resulting in two coefficients each. Also for each vowel, the likelihoods of the MFCC GMMs are computed and converted into posterior probabilities using equal priors. The log posterior probabilities for stress class 0 and 1, along with the normalized duration and the polynomial coefficients for pitch, energy and spectral tilt, are concatenated into a segment-level feature vector for each vowel. Finally, for each of these vectors, the likelihoods of the segment-level GMMs are computed and converted into posterior probabilities using priors learned from data. The final decision on stress level for each syllable is made based on these posterior probabilities using the algorithm described in Fig. 2.

applications. EduSpeak uses a standard GMM hidden Markov model (GMM-HMM) speech recognizer. Recognition is run in forced alignment mode, where the output is constrained to the words in the transcription, using a single forward pass. A 39-dimensional acoustic speech feature is used, which consists of energy and 12 MFCCs, plus their deltas and double deltas. The cepstrum is normalized using cepstral mean subtraction (CMS) with normalization coefficients computed over the entire sentence.

For recognition of the native English data, we used ASR models trained on short utterances from children aged 4–14, with a total of 52,300 utterances from 342 speakers, approximately half male and half female. For recognition of the non-native English data from L1-Japanese children, we used ASR models trained on 47,548 short utterances from 301 native English-speaking children aged 10–14 and adapted to 7119 short utterance from 148 Japanese children on the same age range. In all cases, the data were gender-balanced. The datasets used to train and adapt the ASR models overlap with the datasets described in Section 3.1, which were used for training and evaluating the stress detection systems. This fact might result in slightly optimistic ASR performance, though we believe this bias to be very small given the relatively large amount of speakers used for training the ASR models.

As discussed later on, in future work, we wish to evaluate performance of the proposed stress detection systems when using human-annotated alignments. The difference between this performance and the one obtained with the inaccurate alignments extracted using the ASR models described above would tell us how much of the error in our system's output is due to alignment mistakes.

2.1.2. Log of normalized duration

The duration of the vowel in the syllable is first normalized by dividing it by the mean vowel duration for all syllables of the same type. The syllable type is determined by concatenating two subtypes:

- **next consonant type**: **ufc** (unvoiced following consonant): the consonant after the vowel is unvoiced; or **vfc** (voiced following consonant): the consonant after the vowel is voiced; or **nfc** (no following consonant): no consonant after the vowel (either another vowel follows, or the vowel is the last in the word).
- **pause type**: **nonpp** (non pre-pausal word): the word is not followed by a pause longer than 0.1 s; or **ppls** (pre-pausal word, last syllable): the word is followed by a pause longer than 0.1 s and this vowel is the last in the word; or **ppws** (pre-pausal word, within-word syllable): the word is followed by a pause longer than 0.1 s, and this vowel is not the last in the word.

The duration normalized by syllable type is further normalized by speech rate by dividing it by the mean of the syllable type-normalized duration for all the vowels within the same word. Finally, the logarithm of this normalized value is computed.

2.1.3. Polynomial coefficients of normalized pitch, energy and spectral tilt

Pitch, energy and spectral tilt measurements are extracted at the frame level (every 10 ms) over the full waveform. Pitch (F0) is approximated by the fundamental frequency estimated by our own implementation of the algorithm described by Talkin (1995). Energy (Eg) is approximated by the root mean square value. The spectral tilt (ST) values are computed as the slope of the fast Fourier transform (FFT), extracted over a window of 20 ms that is shifted every 10 ms. The use of spectral tilt was motivated by the findings of Sluijter and Van Heuven (1996). In the following, F0 and Eg refer to the logarithm of the signals extracted

as described above, while ST is not transformed. The exact same processing is done for the F0, Eg and ST signals, as follows:

- Turn the F0, Eg and ST values that correspond to unvoiced frames, as indicated by a missing F0 value, into undefined values. Undefined values are ignored during the computation of the polynomial approximation.
- For each word, find the mean of these signals over the frames corresponding to the vowels. Only defined values are taken into account to compute this mean.
- For each word, subtract the computed mean from the signals only over defined values.
- For each vowel in each word, compute the Legendre polynomial approximation of order 1 for the three signals, which results in 2 coefficients for each signal.

The use of Legendre polynomials was proposed for language identification using prosodic features (Lin and Wang., 2005). This paper can be consulted for details on the computation.

2.1.4. MFCC log posteriors

The same MFCCs used to obtain phone alignments are modeled at the frame level (every 10 ms) over the vowels using one GMM for each stress class. These GMMs were obtained by adaptation to a single GMM trained using samples from all stress classes in the same way as for segmental features (see Section 2.2.1). Given a test utterance, the likelihood of each of these three GMMs is computed for each frame over each vowel. The geometric mean of the likelihoods over all frames in a vowel is computed for each stress class, resulting in three segment-level likelihoods, one for each stress class. These likelihoods are transformed into posterior probabilities using Bayes rule, assuming equal prior probabilities for all stress classes. Finally, the log of the posterior probabilities for stress classes 0 and 1 are used as segment-level features. The posterior probability for class 2 is redundant, given the other two; hence, it is discarded.

2.2. Modeling approaches

Three modeling approaches were compared in our experiments: GMMs, decision trees and neural networks. This section describes the three approaches.

2.2.1. Gaussian mixture modeling

The five types of segmental features – two polynomial coefficients for pitch, two for energy, and two for spectral tilt; plus log normalized duration; plus the log MFCC posterior probabilities for stress classes 0 and 1 – were concatenated into a single feature vector per vowel of size 9. These feature vectors were then modeled with one GMM for each stress class. This modeling was done in two steps. First, a single model for all stress classes was trained. Then, the model was adapted to the data from each stress class. This procedure allowed us to train robust models for even the secondary stress class, for which a very little amount of data is available in comparison to the other two stress classes. The adaptation was done using a maximum a posteriori (MAP) approach commonly used for speaker recognition (Reynolds et al., 2000). This method allows for a regularization parameter, the relevance factor, that controls how much the global means, weights, and covariances should be adapted to the data from each class.

Given a new utterance, we compute the likelihood of the GMM for each of the three stress classes for each vowel. The likelihoods are converted into posterior probabilities using Bayes rule and a set of prior probabilities. These prior probabilities should be computed from data as similar to the test data as possible.

2.2.2. Decision trees

Decision trees have been shown to outperform support vector machines, Naive Bayes and logistic regression by Deshmukh and Verma (2009) for the task of stress detection. Furthermore, decision trees are a standard tool for modeling prosodic features. For this reason, we compared the performance of our proposed GMM approach for segmental feature modeling with that of decision trees. We used CART-style decision trees implemented in the IND toolkit (Buntine, 1992). Deshmukh used C4.5-style decision trees (Duda et al., 2001), but our experiments showed that CART-style significantly outperforms C4.5-style.

To give decision trees a fair chance of outperforming GMMs, we tried two common approaches for improving decision tree performance. We upsampled the minority classes to obtain equal counts for all three classes to allow the trees to describe all classes with equal detail. This approach did not result in a performance improvement in our data; hence, it was not used in our experiments. We also implemented bagging, a technique by which the training data are sampled with replacement N times, thus training a separate decision tree for each sample. The posterior probabilities generated by these trees are then averaged to obtain the final posterior probabilities.

The posterior probabilities generated by the trees were transformed to reflect the desired prior probabilities. As in the case of GMMs, these prior probabilities can be chosen to coincide with those in the native data or the non-native data, or chosen arbitrarily to match the prior probabilities expected during testing of the system. The posterior probabilities for the three classes were transformed by multiplying each of them by the ratio of the desired prior probability for the corresponding class divided by the prior probability seen during training for this class. The resulting posterior probabilities for the three classes were finally normalized to sum to one.

CART-style decision trees use cost-complexity pruning with cross-validation. We modified IND code to accept a file indicating what the subsets for cross-validation should be (rather than determining them randomly, as in the original IND code) and defined the subsets by speaker such

that all samples from a speaker occur in the same subset. This is essential for good performance when bagging or upsampling, since repeated samples occurring in different subsets break the pruning approach and result in large trees that overfit the training data. We use 10 subsets for cross-validation.

Note that, even when decision trees were used for segmental-feature modeling, segment-level MFCC posterior probabilities were still generated using GMMs. This way, in our comparisons, the features modeled by the decision trees and the segmental GMMs are identical.

2.2.3. Neural networks

Neural networks and, in particular, deep neural networks, have had great success in a large variety of problems. Recently, deep neural networks have been shown to outperform GMMs in a few different speech processing tasks. In this work, we explored the use of different multilayer perceptron NN architectures, including deep architectures with several hidden layers.

The NNs had one input node per input feature and one output node for each stress class, which were set to 1 for the stress class corresponding to the sample, and 0 for the other classes. The NNs were trained with a multiclass cross-entropy objective, hyperbolic tangent activation functions in the hidden layers and softmax activations in the output layer. The backpropagation algorithm was used for training the network parameters. The input features were normalized to have zero mean and standard deviation of one before NN modeling, using the statistics obtained on the training data.

The posterior probabilities generated by the NNs were transformed to reflect the desired prior probabilities with a procedure identical to the one used for the decision tree approach. Finally, as also done for decision trees, when using NNs for segmental feature modeling, segment-level MFCC posterior probabilities were still generated using GMMs for consistent comparison of performance with the other two modeling approaches.

2.3. Making decisions

A straightforward way to make decisions based on the posterior probabilities generated by the system is to simply pick the class with the highest posterior probability. This simple procedure may result, depending on the data, in an unacceptably high level of false corrections for a CALL system (see Section 3.2 for a definition of false correction). In general, we wish to control the maximum level of false corrections that is acceptable for the system. For this reason, we used the algorithm depicted in Fig. 2 to reach the final decisions.

The algorithm takes the posterior probabilities generated by the GMM, decision tree, or neural network and the canonical (correct, as indicated by a stress dictionary) stress level for each syllable, c. For each syllable, the posterior probability p_c for the canonical stress level is compared with a threshold t. If the posterior probability p_c is larger than t, the syllable is labeled as having the canonical stress c. Otherwise, the stress class with largest posterior probability for the syllable is chosen. That is, the system will choose to output the canonical stress for a syllable if (1) its posterior probability is larger than a predefined threshold, or (2) its posterior probability is the largest of the three posterior probabilities for the syllable. By varying the threshold, we can then control the level of false corrections. We refer to this as the "benefit of the doubt" algorithm, since the system will only claim an incorrect (non-canonical) stress was produced if the posterior probability for the canonical class is lower than a threshold; otherwise, it gives the canonical label the benefit of the doubt. The threshold t is determined empirically, based on the acceptable level of false corrections for the application.

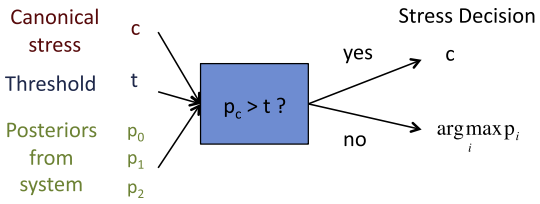

Fig. 2. Proposed decision making algorithm. For each syllable, the posterior probability p_c for the canonical stress level c is compared with a threshold t. If the posterior probability p_c is larger than t, the syllable is labeled as having the canonical stress c. Otherwise, the stress class with largest posterior probability for the syllable is chosen.

3. Experiments

In this section we describe the methods used for the experiments, present statistics on the data and show results for a variety of systems and performance measures.

3.1. Data

Experiments were run on a dataset of L1-Japanese children, aged 10–14, reading English phrases. We call this set the "non-native" set. The children were in the process of learning English and had different levels of proficiency. A set of 959 multisyllabic words was selected from this dataset and labeled by three annotators for stress level. These words came from 668 randomly chosen phrases from 168 distinct speakers from both genders. The chosen speakers were those with a larger number of stress pronunciation errors as judged by an initial quick annotation performed on the data from the full set of 198 speakers, in which stress pronunciation quality was judged at the word level as either correct or incorrect. Multisyllabic words from the remaining 30 "better" speakers were used to compute syllable-type statistics for which we do not require stress labels. These statistics were used to normalize vowel duration for these data.

Annotators were instructed to label each syllable in each selected word from the 168 chosen speakers with a label of

"unstressed" (0), "primary stressed" (1) or "secondary stressed" (2). Annotators were allowed to label more than one syllable with primary or secondary stress. Words for which the number of pronounced syllables did not correspond to the number of syllables in the canonical pronunciation (according to at least one annotator) were discarded. This resulted in 848 words that were labeled by the three annotators and correspond to 1776 syllables; most words were bisyllabic words. These data were only used for testing. They were not used for training or calibration of our systems, except for the experiments in which we adapted models or learned prior probabilities from the non-native data. In all these cases, a cross-validation technique was used to avoid reporting optimistic results.

Unless otherwise stated, results reported in this paper were computed on the set of syllables for which all three annotators agreed on the same stress label, which contains 1240 syllables. The selection was done at the syllable level. Annotators might disagree on the label for some syllables in a word, but not for others. In order to preserve the largest amount of data as possible, we kept syllables for which there is agreement, even if they came from a word containing some other syllable or syllables for which annotators disagree. We call this set the *agreement set*.

A separate dataset of native English-speaking children aged 10–14 was used for training the models and system calibration. We call this set the "native" set. The data consist of read speech from 329 children with a total of 41,022 phrases. All children were native speakers of American English and from the west coast of the United States. All were considered to speak using the "standard" or "accepted" pronunciation of that region. There were approximately equal numbers of boys and girls. Multisyllabic words for which a single stress pronunciation is listed in our lexical stress dictionary were selected. The canonical stress found in the dictionary was then assigned as label for each of these words. We assume that native English speakers pronounced stress as listed in the dictionary in the vast majority of cases for these words. This assumption allowed us to use a large amount of data for training the models without the need for manual annotations. This database contains 74,206 words with a total of 157,888 syllables.

The syllable type statistics used to normalize vowel duration for the native English data were computed on the native data itself. As mentioned above, the statistics used for normalizing duration for the non-native data were computed on a held-out set of non-native L1-Japanese speakers. On preliminary experiments performed on two separate datasets of adult data, we found that the use of statistics from matched datasets gives better performance than the application of the native data statistics to the non-native data. This improvement was also confirmed on the children's dataset in which an increase in error rate of around 1% absolute was observed when the native duration statistics were used on the non-native data, as compared to using statistics computed on the held-out set of non-native speakers. This difference might in part be due to consistent errors performed by the ASR system used to obtain phone-level alignments, which are likely to be different for native and non-native data. The computation of vowel duration statistics by syllable type was the only part of the system that required held-out non-native data. Note that these held-out data used for statistic computation do not require stress labeling.

3.2. Performance metrics

We will use three different performance metrics:

- **Error rate:** The number of samples in which the detected label disagrees with the annotated label divided by the total number of samples times 100. This metric can also be used to compute disagreement between annotators as in Section 3.3.1. It does not involve the use of the canonical stress labels. The accuracy, computed as 100 minus the error rate, is the most standard metric in this field. We chose to report error rates instead, since we believe they give a more intuitive feel of the differences in performance across systems: for example, a change in accuracy from 90% to 92% is less descriptive of the system's improvement than a reduction in error from 10% to 8%, which corresponds to a 20% relative reduction in error rate.
- **False corrections:** The percent of times that the system detects a correctly pronounced syllable as incorrectly pronounced, mistakenly correcting the user. This metric uses the canonical labels as determined by a stress dictionary to decide whether a syllable was correctly or incorrectly pronounced. Since stress labels for natives are always assumed to be the canonical ones, this metric only makes sense for non-native data.
- **Missed corrections:** The percent of times that the system detects an incorrectly pronounced syllable as correctly pronounced, missing a chance to correct the speaker. This metric, like the false corrections, only makes sense for non-native data.

To compute these measures, we used the agreement set for which all three annotators agreed on the label (0, 1 or 2), unless otherwise indicated. When results for stressed versus unstressed are presented – the task we call 02|1 – all syllables annotated as 2 in the agreement set are mapped to 0.

These measures were computed on hard decisions. In order to go from posterior probabilities to decisions, we used two different procedures:

- **Maximum Posterior (maxp):** This is the standard procedure used in the literature when the modeling technique outputs posterior probabilities or likelihoods. The stress class for which the posterior probability is the largest is chosen as the system's output for the syllable. When results on the 02|1 task are presented, the posterior

probability for the unstressed class is computed as the sum of the posterior probabilities for unstressed (0) and secondary stressed (2).

- **False Correction at 5% (fc5):** Using the algorithm described in Section 2.3, we set the threshold on the posterior probabilities to achieve a 5% false correction rate.

For the most important comparisons of two systems, we report the *p*-value obtained with the McNemar matched-pairs significance test.

3.3. Statistics on data

This section presents statistics on the data used for the experiments.

3.3.1. Annotator agreement

The disagreement between annotators, computed in terms of error rate (see Section 3.2 for the definition), is given in Table 1. In the second column, the target task of labeling each syllable with three levels of stress is considered. We find an average disagreement rate of 21% for this task. In the third column, all syllables labeled as 2 have been relabeled as 0. In the stress detection literature, standard practice considers secondary stressed and unstressed syllables as one class. We see that the disagreement is much smaller for this task. The fourth column shows only the disagreement on samples for which at least one of the annotators thought the pronounced stress was correct for the task of telling unstressed versus stressed. Pronounced stress is considered correct if the assigned label coincides with the canonical one; otherwise, it is incorrect. For the purpose of this computation, canonical 2s have also been mapped to 0s. The last column shows only the disagreement on samples for which at least one of the annotators thought the pronounced stress was incorrect.

We can see that the disagreement is much smaller on syllables that were labeled as correctly pronounced by at least one of the annotators. This indicates that agreement is easier for syllables that are correctly, rather than incorrectly, pronounced. Correctly pronounced syllables have a more familiar sound; incorrectly pronounced syllable can be uttered in ways that are not standard for native speakers and, hence, harder to label as either correct or incorrect.

Table 1
Percent of disagreement between annotators when all three stress classes are considered separately (0|1|2) and when class 2 is merged with class 0 (02|1). The last two columns correspond to the disagreement on the 02|1 task when only samples for which at least one annotator labeled the syllable as correct or incorrect are selected.

Annotators	0|1|2	02|1	02|1 cor	02|1 inc
A1 versus A2	21.3	16.5	19.3	52.8
A1 versus A3	18.2	10.0	11.3	45.7
A2 versus A3	23.6	16.7	19.4	54.5
Average	21.0	14.4	16.7	51.0

3.3.2. Statistics on stress labels

Table 2 shows the proportion of 0s, 1s, and 2s in the native data and the non-native data on the agreement set. The distribution of labels on the non-native data is significantly different from that in the native data, with a large increase in the proportion of stressed syllables. Of the 191 words for which the three annotators agreed were incorrectly pronounced, 35% of them were labeled as having at least two primary stressed syllables.

Finally, Table 3 shows the confusion matrix of canonical stress versus labeled stress for the non-native data. This table shows that most primary stressed syllables are pronounced correctly. On the other hand, around half of the unstressed syllables are pronounced with primary or secondary stress. Secondary stressed syllables are also stressed with more stress than they should in 32% of the cases. Clearly, Japanese children tend to overstress syllables when speaking English, even when this results in more than one syllable in a word having primary stress.

Overall, the agreement set contains 1240 syllables, 78% of the syllables labeled as correctly pronounced (that is, labeled with the canonical stress label) and 22% labeled as incorrectly pronounced. While the percent of syllables labeled as correct for each individual annotator is between 63% and 67%, for the agreement set it is 78%, coinciding with the observation made for Table 1 that the agreement is higher for syllables that were labeled as correct by at least one annotator.

Note that these statistics tell us that the agreement set is somewhat biased with respect to the prior probabilities of the stress levels. That is, our evaluation set contains a larger percent of correctly pronounced syllables that would be found in the original set. This will also affect the systems that use prior probabilities computed on the non-native data. Unfortunately, this is an inherent issue in this data, which cannot be easily solved. One possible way to avoid this bias would be to ask annotators to discuss every single

Table 2
Percent of syllables labeled with each stress class for natives and non-natives.

Dataset	%0	%1	%2
Natives	48.3	47.2	4.5
Non-natives	21.9	67.1	11.0

Table 3
Percent of syllables labeled with each stress class on the agreement set for non-native speakers for each canonical stress class (diagonal highlighted in bold). The total amount of syllable with each canonical stress class is shown in the last column.

Can	Lab			Count
	0	2	1	
0	**54.00**	17.66	28.34	487
2	3.23	**64.52**	32.26	31
1	1.11	4.16	**94.74**	722

case of disagreement until they come to an agreement, assigning a single consensus label to each syllable. This would be a difficult and costly process. Section 3.4.5 presents an analysis of the effect of this selection by computing results on the set of syllables for which 2 of the 3 annotators agreed on the same stress level.

3.4. System performance

Results in this section were obtained using (1) CART-style decision trees; (2) neural networks; (3) a set of "big" models of 2048 Gaussian components for MFCC modeling and a 512-component GMM for segmental feature modeling; and (4) a set of "small" models of sizes 256 and 64, respectively. All models were trained on the native data described above. The big GMMs were tuned to minimize error rate on 10-fold cross-validation experiments on native data. While these GMM sizes might seem very large compared to those used in ASR and other tasks, they are comparable to the sizes used in state-of-the-art speaker recognition (see, for example, (Ferrer et al., 2013)), in which, like this task, many phones are modeled with a single GMM. Nevertheless, as we will see, these big sizes were unnecessary for optimal performance on non-native data. We show results on the smaller models for comparison.

3.4.1. System comparison

Table 4 shows results for several different systems and setups for the task of classifying stress into three levels: 0 (unstressed), 1 (primary stress) or 2 (secondary stress).

Table 4
Error rate for the maximum posterior probability decisions (maxp) for native and non-native children's data on two tasks: 0|1|2 and 02|1. For the non-native data and the 0|1|2 task, we also show the miss rate when setting the threshold for a false correction rate of 5% (fc5). We show results on four sets of systems: decision trees (DT), neural networks (NN), and big and small GMMs for MFCC and segmental modeling. For non-natives we show results on the full set of words; on a subset of words that was labeled as correctly pronounced (cor words); and on a subset of words that was labeled as incorrectly pronounced (inc words). The GMM systems are: **native p, sep trn**, in which native prior probabilities are used for posterior probability computation and class-dependent GMMs are trained independently; **native p**, in which native prior probabilities are used and class-dependent GMMs are obtained through adaptation to a class-independent model; **non-nat p**, a system identical to the previous one, but in which non-native prior probabilities are used for posterior probability computation instead of native prior probabilities; and **non-nat p, adapt to non-nat**, a system identical to the previous one, but in which an additional step of adaptation to non-native data is done on the class-dependent GMMs. For the DT and NN systems, we compare the use of native and non-native prior probabilities. For DT, we also show results without using the bagging approach. The bolded numbers correspond to the best all-word performance for each dataset and each task. For the first two columns we show significance level between the system corresponding to the line and the one in the previous line within the same block.

Task	System	Setup	Maxp %Error				fc5 %Miss
			Native	Non-native			Non-native
				All words	cor words	inc words	All words
0|1|2	DT	Native p, no bagging	14.1	23.6	14.2	44.4	73.3
		Native p	13.7***	23.2	12.5	45.4	64.5
		Non-nat p		21.0**	16.8	38.9	52.0
	NN	Native p	14.4	22.6	12.7	44.4	64.8
		Non-nat p		21.0	20.1	36.7	52.4
	GMM Small	Native p, sep trn	13.8	24.3	14.4	44.9	70.0
		Native p	13.3***	22.9*	11.8	46.4	63.7
		Non-nat p		20.3**	18.6	35.7	52.8
		Non-nat p, adapt to non-nat		**20.2**	19.0	35.7	52.0
	GMM Big	Native p, sep trn	12.5	25.7	14.4	47.6	70.7
		Native p	**11.5****	22.5***	11.6	43.6	65.9
		Non-nat p		21.3	18.6	37.9	54.2
		Non-nat p, adapt to non-nat		20.3	17.9	36.4	**48.4**
02|1	DT	Native p	11.8	16.8	10.3	28.4	
		Non-nat p		16.6	12.9	28.9	
	NN	Native p	12.5	16.5	9.6	28.4	
		Non-nat p		15.4	13.8	25.9	
	GMM Small	Native p	11.3	16.2	8.5	29.2	
		Non-nat p		15.0	12.5	25.7	
		Non-nat p, adapt to non-nat		14.6	12.9	24.9	
	GMM Big	Native p	**9.8**	16.8	8.8	28.9	
		Non-nat p		15.0**	10.9	26.4	
		Non-nat p, adapt to non-nat		**14.5**	11.4	25.7	

* Indicates a p-value smaller than 0.05, respectively.
** Indicates a p-value smaller than 0.01, respectively.
*** Indicates a p-value smaller than 0.001, respectively.
An absent symbol indicates a p-value larger than 0.05.

We call this task 0|1|2. We also show results for the task of classifying syllables into unstressed (0 or 2) or stressed (1) for a subset of the system setups. We call this task 02|1.

Different setups for each of the models (decision trees, neural networks, small GMMs and big GMMs) are compared. In all cases, systems using prior probabilities computed on native data ("native p") and non-native data ("non-nat p") are compared. For the GMMs, a baseline system in which GMMs for the different classes were trained separately without the adaptation technique described in Section 2.2.1 is presented ("sep trn"). The three GMM systems without the "sep trn" label use the proposed adaptation technique with a relevance factor of 0, which gave optimal performance on native data. A fourth system is presented for the GMM approach, in which the class-dependent models were adapted to the non-native data ("adapt to non-nat") in a second step of adaptation. This was done with the MAP approach described in Section 2.2.1. In this case, though, given the small amount of non-native data available, only means and weights were adapted.

Since MFCC GMMs were trained on the native data which were then used again to learn the segmental models, we used a 10-fold cross-validation procedure to create the log-posterior probabilities from the MFCCs. The MFCC GMM models were trained on 9/10th of the data and used to create the MFCC log-posterior probabilities for the remaining 1/10th of the data; the sets were rotated until MFCC log-posterior probabilities had been computed for all data.

Table 4 shows error rate results for native and non-native speakers using the maximum posterior probability algorithm for making decisions. For the non-nat p systems, we only show results on non-native data, since these systems are only meant to optimize performance on those data. For the native results we did 10-fold cross-validation, training the system on 9/10th of the speakers and testing it on the held-out 1/10th of the speakers; finally, we collected posterior probabilities from all subsets to compute the shown performance. The prior probabilities used for posterior probability computation were computed on the subsets used for training and applied on the test subset.

When using non-native prior probabilities and when doing adaptation to non-native data, we used the same 10-fold cross-validation approach. The relevance factor used for adaptation to non-native data, on the other hand, was selected to optimize the performance on the full set, which means these results are somewhat optimistic. The optimal relevance factor was 80 for the small models and 5 for the big models. For non-native data, Table 4 shows results for two additional subsets of data consisting of only the words that were labeled as correctly or incorrectly pronounced by the three annotators. The number of words labeled as correct and incorrect is 220 and 191, respectively.

Finally, for non-native data we also show the miss rate obtained when setting the posterior probability thresholds at a 5% false correction rate. The threshold was trained on all the test data to ensure that exactly a 5% false correction rate was achieved. Thresholds could be determined using the cross-validation approach, but this would result in biased thresholds when the data are also used for adaptation of the models. Nevertheless, we found that, except for this last system in which thresholds cannot be determined through cross-validation, in all other cases, miss rates and false correction rates with thresholds determined through cross-validation or on the full test data were very close. The absolute difference in miss rates and false correction rates when using the thresholds obtained with cross-validation or on the full test data was smaller than 1% and 0.1%, respectively.

Results for the DT and the NN approaches are also shown in Table 4. For DTs we show results with and without bagging. For NNs, we show results using two hidden layers with 400 nodes each. This architecture was optimal on the native data and also on the non-native data when using the non-native priors. We can see that DTs and NNs are somewhat competitive with the small GMM systems for the 0|1|2 task, but significantly worse than the big GMM systems for native speakers. This suggests that GMMs are better models than the other two when matched training data are available. Hence, if a large amount of non-native data was available to directly train models for non-native speech, the GMM approach would probably be preferable. Furthermore, for the 02|1 task, even the small GMMs significantly outperformed decision trees.

For the GMM systems, we can see that the proposed adaptation technique for training class-dependent models gives significant reductions in error rate, especially on non-native data and when using big models. The latter is expected, since training big models is more prone to overfitting than training small models. Using prior probabilities calculated from non-native data gives modest gains on these data. These gains are also expected given the big difference in prior probabilities between native and non-native data shown in Table 2. Finally, adapting the segmental GMM parameters to the non-native data gives some further gain, though not a statistically significant one. Note that the last two systems require some amount of labeled non-native data, while the first two only use native data for model creation.

Overall, we see that big models give significantly better performance than small models for native data but not for non-native data. This implies that the details modeled by the big system are too specific to the native data and do not generalize well to non-native data.

We also see that the performance on non-native speakers is significantly worse than on native speakers. This degradation comes from the incorrectly pronounced words, since performance on correctly pronounced words when using native prior probabilities is comparable to the one obtained on native speakers. This suggests that the system has more difficulty classifying mispronounced words. This can be due to both issues with the ASR alignments (although even correctly stressed words might be

misaligned due to phonetic mispronunciations) and to the fact that incorrectly pronounced stress might be labeled as such because it was pronounced in a non-native manner with an unusual combination of duration, energy and pitch patterns. These patterns would not have been seen under any stress class in the native data. This could suggest that using non-native data for training or adaptation should improve performance on these words. Nevertheless, the table shows that adapting models to non-native data does not bring large improvements. Furthermore, training the models directly on this data results in significant degradations for all modeling approaches, NN, DT and GMM (results not shown). We believe that the lack of a significant gain from training models on or adapting models to non-native data is due both to the small amount of data available and to the high disagreement between annotators on incorrectly pronounced syllables that results in very noisy data.

As expected, results on the simpler task of classifying stressed versus unstressed syllables were significantly better than those on the 3-way classification task. Note that these results are only approximations, since they were obtained by mapping any syllable annotated as a 2 to a 0. A label of 0 might not have been the preferred choice for all these syllables if annotators had been forced to label each syllable as either 0 or 1. Hence, these numbers should be seen as an approximation for the accuracy we would obtain with this system on a database labeled with only unstressed and primary stressed syllables. This approximation is likely to be pessimistic since, while new errors would appear if some of the 2s were labeled as 1 instead of 0, more errors would probably disappear, given that the system is correct more times than not.

Finally, conclusions drawn from miss rates for fc5 decisions are consistent with those drawn from error rates for maximum posterior probability decisions.

3.4.2. Feature design decisions

Many choices were made during feature design. The most important ones were: (1) the normalization procedure for the duration feature (whether to do syllable-type normalization, or to do speech rate normalization); (2) the use of the logarithm of normalized duration instead of raw normalized duration; (3) the normalization procedure for the pitch, energy, and spectral tilt features (whether to normalize the signals with the frame-level mean over the vowels before polynomial approximation occurs, or to normalize the 0th-order Legendre polynomial coefficient with its mean over the vowels); (4) the polynomial degree for modeling these signals (values from 0 to 5 were tried); (5) the use of geometric mean instead of arithmetic mean to compute the MFCC segmental likelihoods; (6) the use of uniform prior probabilities to convert the MFCC likelihoods into posterior probabilities; and (7) the use of logarithm of the MFCC posterior probabilities, rather than the raw posterior probabilities, as segmental features. Each of these decisions was made by running comparison results on the native data and, in some cases, on a separate dataset of adult Japanese speakers where some preliminary experiments were run.

Decisions 2, 3, 5, 6, and 7 did not make a significant difference in performance that would warrant a detailed comparison in this paper. The polynomial degree of 1 (decision 4) gives around 5% relative improvement in error rate with respect to using an order of 0. Higher orders do not lead to further gains; they also make the system more complex. The duration normalization procedure (decision 1) also has a significant effect in performance. The absence of any type of normalization increases the error rate on native speakers by around 12% relative to the use of both types of normalization. The use of speech rate normalization without syllable-type normalization increases the error rate by around 7%; conversely, the use of syllable-type normalization without speech rate normalization increases the error rate by around 10%. For the non-native data, a lack of syllable-type normalization results in a 9% degradation, and a lack of speech rate normalization does not result in a significant difference in results. Feature selection results will be shown in Section 3.4.3 to justify the inclusion of all five types of features in the model.

Different procedures were also tried for the integration of spectral and segmental information. The frame-level pitch, energy, and spectral tilt signals and their deltas were appended to the MFCC features and the duration feature was replicated and appended to the resulting vector for all frames in a vowel. The problem with this approach is that unvoiced frames (which happen sometimes even within vowels) have meaningless pitch values. As another alternative, pitch, energy, spectral tilt, and duration segmental features computed as described in Section 2.1 were replicated for all frames in a vowel and appended to MFCCs. In these two options, a single set of frame-level GMMs were trained and used to generate segment-level posterior probabilities as explained in Section 2.1.4. As a final alternative, the 13 static MFCCs were taken as separate signals and modeled using polynomial approximations as for the other frame-level signals. The resulting segment-level coefficients were appended to the other segment-level features and modeled with GMMs. All of these approaches proved to be significantly worse than the approach described in Section 2.

3.4.3. Feature selection results

The proposed system uses five types of features based on pitch, energy, spectral tilt, duration and MFCC information. Fig. 3 shows results using all possible combinations of feature types. The system setup is kept identical to the one used for the small "native p" system in Table 4. The results shown correspond to the error rate for natives and non-natives for the maxp decisions.

For natives, we see that the best result for each value of N is always better than the best result for $N-1$. This is true even for the 5-feature result (13.26%) which is better than the best 4-feature result (13.58%) at a significance level

smaller than 0.001. From this, we can conclude that all five feature types are needed to achieve the best performance. Interestingly, the best N-feature combination always includes the best $N-1$-feature combination: the single best feature type is MFCC; the best 2-feature combination adds duration to that; the best 3-feature combination adds energy to those two; the best 4-feature combination adds pitch; and, finally, the 5-feature combination adds spectral tilt, still giving a statistically significant gain. Given this, one could order the features in terms of importance for stress classification on native speakers as follows: (1) MFCCs, (2) duration, (3) energy, (4) pitch, and (5) spectral tilt.

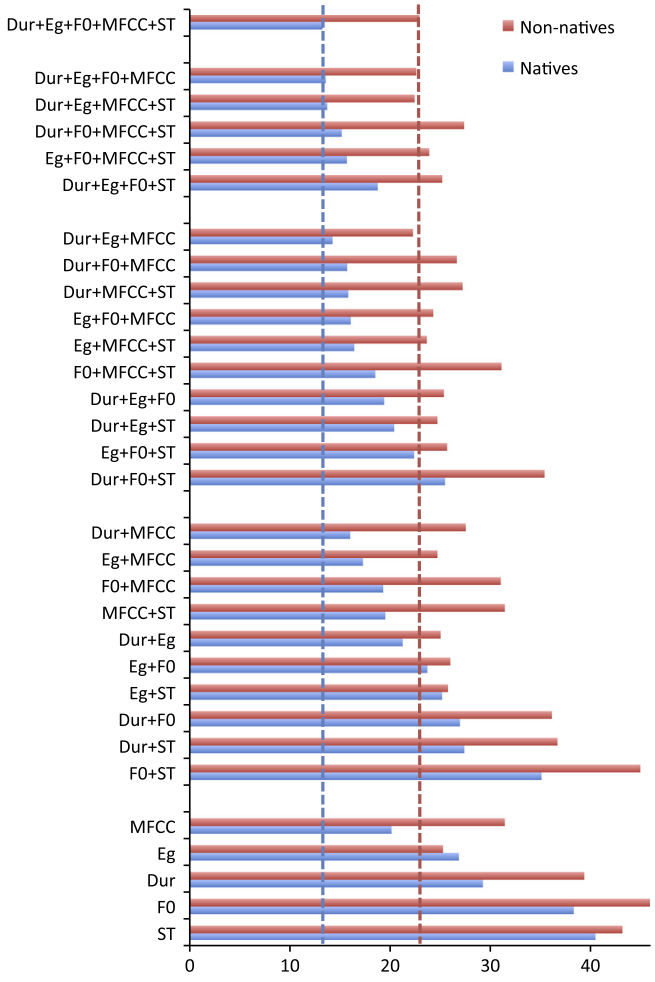

Fig. 3. Native and non-native error rates for all possible feature combinations. For each N, systems are sorted based on their native performance. The two vertical lines indicates the all-feature performance for natives and non-natives.

Feature types can also be ordered based on the performance loss that occurs when that single type is taken out of the combination (that is, looking at the combinations of four feature types). Taking out MFCCs gives the largest loss, followed by duration, energy, pitch and spectral tilt, coinciding exactly with the order obtained by the cumulative nature of the N-feature results.

The order of importance of feature types is slightly different for non-native data. The same cumulative nature of the best N-feature combination is observed for non-natives, except that the best single feature is energy, the best 2-feature combination adds MFCCs to the energy, and the best 3-feature combination adds duration. In this case, pitch and spectral tilt give a non-significant degradation when added. As with the native data, this same order can be obtained by looking at the 4-feature results. Nevertheless, unlike for the native data, only the performance loss when discarding energy and MFCCs are statistically significant, with p-values smaller than 0.005.

We chose to present feature selection results using native prior probabilities because we believe that this approach gives a more direct assessment of the usefulness of the feature itself. The non-native prior probabilities bias all systems toward detecting more primary stressed syllables, thus washing out differences across feature types. Despite this, using non-native prior probabilities for the non-native data still results in energy being the single best feature, now followed by duration and then MFCCs, with all three of them giving gains over the all-feature result.

3.4.4. Controlling false correction rate

As described in Section 2.3 we propose an approach for decision making that allows us to control the amount of false corrections made to the user. While we show fc5 results in Section 3.4.1, in this section we give a more detailed comparison of the results obtained with the maxp approach and the fc5 approach. Table 5 compares error rates, false correction rates and miss rates for the maximum posterior probability decisions and the decisions based on the proposed algorithm, with the aim of achieving no more than 5% false correction. Results correspond to the big "non-nat p, adapt to non-nat" system from Table 4. The threshold for the fc5 decisions was determined on the full

Table 5
Error rate, false correction rate and miss rate for non-native children's data on task 0|1|2 using the big "non-nat p, adapt to non-nat" system from Table 4. We show results on the full set of words (all words); on a subset of words that was labeled as correctly pronounced (cor words); and on a subset of words that was labeled as incorrectly pronounced (inc words). We compare two algorithms for decision making: maximum posterior probability and the proposed algorithm, fc5. In the case of correct words, the error rate coincides with the false correction rate and the miss rate is 0, since there are no errors that can be missed.

Decision algorithm	All words			Corr words			inc words		
	%error	%fc	%miss	%error	%fc	%miss	%error	%fc	%miss
maxp	20.3	11.9	28.6	17.9	17.9	0.0	36.4	8.2	31.7
fc5	17.7	5.1	48.3	7.9	7.9	0.0	41.4	3.5	53.0

non-native data, as in Section 3.4.1. Native results are not shown here because false correction and misses are only meaningful when pronunciation errors are made.

Giving the canonical stress label the benefit of the doubt results in the correct labeling of more of the correctly pronounced syllables. This is clearly shown in the results for the correct words, which see a decrease in error rate of 56%. On the other hand, incorrectly pronounced words see an increase in error rate, though a rather small one. While the fc5 algorithm achieves the desired goal of controlling the false correction rate, it also results in a much larger rate of misses than the maximum posterior probability algorithm: around half of the incorrect syllables (48.3%) go unnoticed by the system, compared to less than one third (28.6%) for the maximum posterior probability algorithm. Nevertheless, we believe this type of error is of much less importance than the false correction error, and we prefer to work at this operating point.

Interestingly, we see that the proposed algorithm, besides keeping the false correction rate at the desired 5% value, also decreases the overall error rate around 13% relative. That is, the proposed algorithm helps approximate the decisions made by the annotators better than the standard algorithm. This simply reflects the fact that 78% of the syllables are labeled as correctly pronounced (all the syllables in the correctly pronounced words and around half of the syllables in the incorrectly pronounced words are correctly pronounced) and these are the samples for which the proposed algorithm is designed to work better than the maximum posterior probability algorithm.

3.4.5. Analysis on syllables with annotator disagreement

As mentioned in Section 3.1 the full set of non-native data contains 1776 syllables, out of which only 1240 were given the same label by all three annotators. All results above are presented on this agreement set. The remaining 536 syllables can be considered gray-area cases: syllables which are hard to label consistently because they contain some combination of characteristics that confuse the annotators. We would like to know how our system performs on those syllables.

To this end, we ran our best system without adaptation ("GMM big, non-nat p") on all syllables for which at least two annotators agreed on the same label. This set contains 1727 syllables. We reran the system using the non-native priors computed on this set (using a 10-fold cross-validation procedure as described in Section 3.4.1). This change in priors did not change the performance on the agreement set, where all three annotators agreed, indicating that the system is robust to small changes in the priors.

On the remaining 487 syllables for which exactly two annotators agreed on the same label, we find that the system's accuracy with the fc5 decisions is greatly degraded, reaching a 55% error rate (compared to 18.5% on the agreement set). Of those 55% samples that are not labeled with the stress level given by the majority of annotators, 42% are labeled with the stress level given by the annotator that disagreed with the other two. Finally, from the remaining mislabeled 58% of the 55%, 90% are labeled with the canonical stress.

In summary, only 3% of the samples for which exactly two annotators agreed on the same label are labeled with a stress level that was not chosen by one of the annotators and is not the canonical stress level for that syllable. We believe this indicates that, whatever the system is doing for these ambiguous syllables will not be bothersome to the user, since, if a mistake is pointed out, it is likely to be one (at least according to some annotator) and, otherwise, the user is given the benefit of the doubt.

3.4.6. Agreement with individual annotators on non-native data

We computed the error rate of the system's output using the labels from each individual annotator on the full set of labeled data (as opposed to the set in which the three agree) as a measure of the disagreement between an automatic annotator (the system) and a human annotator. The error rate of the big "non-nat p" system (with prior probabilities computed on the agreement set) using fc5 decisions on the 0|1|2 task ranges from 30.2% to 33.22% for the different annotators. Comparing the numbers in Table 1, we see that the system is still not at the level of human performance. We believe this is very likely due in the most part to the error-prone alignments provided by the ASR system. While automatic alignments are always error-prone, this is especially true for non-native speech in our system, since it was trained with more native than non-native data. Deshmukh and Verma (2009) showed a significant effect of ASR quality in their stress classification system, going from 79% accuracy to 89.9% when using an improved ASR system. In agreement with these results, a careful review of our system's output by a highly trained phonetician indicated that a large proportion of the errors made by the system occurred in syllables in which automatic alignments were significantly off. This suggests that a key focus for improving stress classification performance should be the improvement of ASR performance.

3.4.7. Comparison of results with previous work

Deshmukh and Verma (2009) show 81.9–90.9% accuracy on their per-syllable experiments on Indian speakers of English when using different modeling methods. The best performing system uses nucleus-dependent modeling. Their results are computed only on words that were labeled as correctly pronounced, since they only label at the word level for correctness; transferal of labels to the syllables can be done only for correctly pronounced words. Words are considered correctly pronounced when their three annotators agreed on it. These results can then only be compared to the "corr words" column in Table 4 for the 02|1 task, since they do not label secondary stress. We see that our system performs similarly to their best configuration with an accuracy of 91.5% (8.5% error rate) when native prior probabilities are used. A significant improvement to the system

presented by Deshmukh and Verma (2009) is shown by Doddala et al. (2011) where information about phonetic context as well as the nucleus identity is included in their models, reaching accuracies of around 96% on bilingual Spanish/English children including only correctly stressed words. As mentioned above, our experiments showed no gains from doing nucleus-dependent modeling or normalization. We believe this is due to the high rate of phonetic mispronunciations present in our database, even when stress is pronounced correctly.

Tepperman and Narayanan (2005) show accuracies between 82.6% and 85.6% at the syllable level for Italian and German learners of English. Their labels enforce a single primary stress per word and no secondary stress. Results for this simplified task can be compared to our results on non-native speakers for the 02|1 task in Table 4, corresponding to 85.5% accuracy (14.5% error rate), except for the fact that their alignments are not automatic but manual, which should result in a significant advantage.

Li et al. (2013) present, to our knowledge, the only stress classification system that attempts to detect three levels of stress (primary and secondary stressed, and unstressed). They use deep belief networks trained on prosodic features from a large amount of matched non-native data. Their system reaches an accuracy of 80% on the three-way classification task, a performance comparable to ours for this same task, except that they only test on 3-syllable words while we test mostly on 2-syllable words.

The only other work we have found that uses native English data to train models that are then applied to non-native English data is presented by Chen and Wang (2010). They show a large degradation in performance of around 12% absolute, when testing on non-native data from Chinese learners of English compared to testing on native English data, with a word-level accuracy on non-natives of around 77%. Since decisions are made at the word-level, these results are not comparable to ours. Nevertheless, for 2-syllable words, the word-level accuracy should coincide with the syllable-level accuracy if a single stressed syllable is enforced as is the case in their work. Their reported result for 2-syllable words on non-native data is 80%. This result can be compared to our non-native results for the 02|1 task in Table 4 which is significantly better (85.5% accuracy).

4. Conclusions

We propose a system for lexical stress classification at the syllable level that uses both prosodic (pitch, energy and duration) and spectral (tilt and MFCC) features. Pitch, energy and spectral tilt features are first extracted at the frame level and converted to syllable-level features using Legendre polynomial approximations. MFCCs, also extracted at the frame level, are converted to syllable-level features by computing log-posterior probabilities, given a set of class-dependent GMMs. All syllable-level features are concatenated into a single vector and modeled using one GMM for each stress class.

We test the proposed system and compare different setups on a database of L1-Japanese children and a database of L1-English children. In both cases the children read English phrases. Our algorithm results in an error rate of around 11% on L1-English data and around 20% on L1-Japanese data. We show that all features, both spectral and prosodic, are necessary for the achievement of optimal performance on the L1-English data, with the MFCC log-posterior probability features being the single best set of features, followed by duration, energy, pitch and finally, spectral tilt features. For L1-Japanese speakers, energy, MFCC log-posterior probabilities and duration are the most important features.

Given our review of results in the literature and the fact that our proposed method outperforms the more standard decision tree and neural network modeling approaches, we believe our system is competitive in terms of performance. Furthermore, it provides more detailed information than most systems in the literature, allowing for multiple stressed syllables in a word and giving syllable-level feedback with three levels of stress. Finally, one of our proposed systems, the one we call “native p”, does not require labeled data from speakers with the same L1 as the test data. This makes it cheap to train (requiring only L1-English data for which labels can be automatically derived) and portable to any population of English learners. We have shown that this system performs only around 10% worse in terms of error rate relative to a system that takes advantage of matched non-native L1-Japanese data for prior probability computation and model adaptation.

Finally, since the system was developed as a pedagogical tool for language learners, we propose a new method for decision-making based on posterior probabilities of stress classes that allows developers to adjust the system’s operating point to a certain maximum level of false corrections. We believe this kind of error is the most bothersome to a learner and should be minimized, even at the cost of increased missed corrections. This approach is not only able to control the false correction rate but also results in a reduction in overall error rate of around 13% relative to using the maximum posterior probability decisions that are standard in the literature.

Acknowledgments

This work was funded by SRI International. We wish to thank the phoneticians, Alan Mishler, Rebecca Hanson and Timothy Arbisi-Kelm for their hard work in labeling the Japanese children’s data.

References

Ananthakrishnan, S., Narayanan, S., 2005. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In: Proc. ICASSP, Philadelphia.

Buntine, W., 1992. Learning classification trees. Stat. Comput. 2 (2), 63–73.

Chen, L.-Y., Jang, J.-S., 2012. Stress detection of English words for a CAPT system using word-length dependent GMM-based Bayesian classifiers. Interdisc. Inform. Sci. 18 (2), 65–70.

Chen, J.Y., Wang, L., 2010. Automatic lexical stress detection for Chinese learners' of English. In: 2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2010.

Deshmukh, O.D., Verma, A., 2009. Nucleus-level clustering for word-independent syllable stress classification. Speech Commun. 51 (12).

Doddala, H., Deshmukh, O.D., Verma, A., 2011. Role of nucleus based context in word-independent syllable stress classification. In: Proc. ICASSP, Prague.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. Wiley.

Ferrer, L., McLaren, M., Scheffer, N., Lei, Y., Graciarena, M., Mitra, V., 2013. A noise-robust system for NIST 2012 speaker recognition evaluation. In: Proc. Interspeech, Lyon, France.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., Cesari, F., 2000. The SRI EduSpeak™ system: recognition and pronunciation scoring for language learning. In: Proceedings of InSTILL 2000.

Franco, H., Bratt, H., Rossier, R., Gadde, V.R., Shriberg, E., Abrash, V., Precoda, K., 2010. EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. Lang. Test. 27 (3), 401–418.

Lai, M., Chen, Y., Chu, M., Zhao, Y., Hu, F., 2006. A hierarchical approach to automatic stress detection in English sentences. In: Proc. ICASSP, Toulouse.

Li, C., Liu, J., Xia, S., 2007. English sentence stress detection system based on HMM framework. Appl. Math. Comput. 185 (2).

Li, K., Qian, X., Kang, S., Meng, H., 2013. Lexical Stress Detection for L2 English Speech Using Deep Belief Networks.

Lin, C., Wang, H., 2005. Language identification using pitch contour information. In: Proc. ICASSP, vol. 1. Philadelphia, pp. 601–604.

Oxman, E., Golshtein, E., 2012. Detection of lexical stress using an iterative feature normalization method. In: Afeka-AVIOS Speech Processing Conference 2012.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10, 19–41.

Sluijter, A., Van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. J. Acoust. Soc. Am. 100.

Talkin, D., 1995. Robust Algorithm for Pitch Tracking. Elsevier Science.

Tepperman, J., Narayanan, S., 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In: Proc. ICASSP, Philadelphia.

Verma, A., Lal, K.l., Lo, Y.Y., Basak, J., 2006. Word independent model for syllable stress evaluation. In: Proc. ICASSP, Toulouse.

Zhao, J., Yuan, H., Liu, J., Xia, S., 2011. Automatic lexical stress detection using acoustic features for computer assisted language learning. In: Proc. APSIPA ASC.

Zhu, Y., Liu, J., Liu, R., 2003. Automatic lexical stress detection for english learning. Proc. 2003 International Conference on Natural Language Processing and Knowledge Engineering. IEEE.