WORD INDEPENDENT MODEL FOR SYLLABLE STRESS EVALUATION

Ashish Verma, Kunal Lal[1]*, Yuen Yee Lo, Jayanta Basak*

IBM India Research Lab, New Delhi, India
[1] Electrical Engineering Department, Indian Institute of Technology, New Delhi, India

ABSTRACT

Analyzing syllable stress in spoken English has been an area of research for a long time. In this paper, we analyze the performance of a novel method for evaluating syllable stress in spoken English. Specifically, we study the problem of determining if a word is spoken with the correct syllable stress pattern. The proposed method uses generalized models for stressed and unstressed syllables to analyze the constituent syllables of a word and determines if the word is spoken correctly. The performance of the proposed method is reported in terms of classification results on human labeled word utterances and it is compared with that of the word-dependent models using various classifiers.

1. INTRODUCTION

Lexical stress plays an important role in speaking and understanding English language. The meaning of a word in English can completely change depending upon which of its constituent syllables is stressed, *e.g.*, 'content', 'project', 'address', etc. The part of speech for a word can also depend upon the syllable stress of the word. Therefore, it is very important to analyze syllable stress in any English language learning application. Syllable stress is also very important in the context of a text-to-speech synthesis system to produce intelligible and natural sounding speech. A labeled speech corpus is often required to train the prosody models of the TTS system which are later used to predict the stress levels in a test utterance. In this paper, we focus on the problem of identifying incorrect lexical stress in English words spoken by Indian speakers. Since many local languages in India, *e.g.*, Hindi and Bengali, don't have word specific syllabic stress, Indian speakers often mispronounce English words from syllable stress viewpoint. We propose a method to determine if a word is correctly stressed at the syllabic level using generalized models of stressed and unstressed syllables.

Several studies have been conducted in the literature to automatically identify stressed and unstressed syllables in a speech utterance [1, 2, 3, 4, 5]. In English, polysyllabic words have one syllable with primary lexical stress and rest of the syllables are either unstressed or may have secondary stress. For simplicity, secondary stressed syllables are often considered the same as the unstressed syllables in these studies as well as in this paper. Most of the studies on syllable stress use three basic acoustic features for identifying stress, *viz.*, fundamental frequency, energy and duration. All the other features are generally derived from these three basic features. The problem of identifying stressed syllable becomes difficult because the nature of variation of these parameters across stressed and unstressed syllables is not uniform and depends upon the syllable itself, the word containing the syllable and the speaker. For example, a syllable containing short vowel will not show large variations in duration when it is stressed as compared to a syllable containing long vowel. Moreover, different speakers use different prosodic features to stress a syllable. In [1], the role of prosodic features, duration, amplitude and fundamental frequency is investigated in identifying stressed syllables in spontaneous speech. It has been concluded that duration and amplitude are more important in identifying stress as compared to fundamental frequency. Accuracies of three different classifiers, *viz.*, neural networks, Markov chains and rule based classifier, to classify stressed and unstressed syllables have been studied in [2]. A syllable stress classifier to study sentence level stress has been described in [3]. Mel Frequency Cepstral Coefficients (MFCC) have also been used in addition to the fundamental frequency and amplitude as acoustic features in [3]. There also have been some attempts to classify various stress conditions, such as, anger, clear, fast, loud [6]. Most of these studies focus on classifying syllables in a speech utterance into stressed and unstressed categories. Only a few attempts have been made to classify a word into correct and incorrect categories from syllable stress point of view [4]. We address this latter issue in this paper and report the performance of various classifiers.

Rest of the paper is organized as follows. Section 2 describes the extraction of various acoustic feature from the recorded word utterances. Word dependent models for syllable stress are discussed in Section 3. Section 4 describes the motivation and mathematical formulation of word independent models used in this paper. The experiments performed to evaluate different approaches and the corresponding results are described in Section 5. Section 6 draws some conclusions from the paper.

2. FEATURE EXTRACTION

The process of acoustic feature extraction and training of word dependent/independent models is schematically shown in Fig. 1. For each word, several utterances are recorded from different speakers. The utterances are time aligned with the corresponding phonetic spelling of the word using the acoustic models and the lexicon dictionary of a speech recognition system. A phone-to-syllable mapping for the word is then applied to get syllable level time alignment of the utterance. Various acoustic features, as described below, are then computed for each of the constituent syllables.

We have used eight acoustic features to build the stress models which are shown in Table 1. Average fundamental frequency, average energy, and duration of the syllable are three main prosodic features used in this paper. We normalize these three main features with the corresponding values over the whole word utterance to remove any speaker dependent variation. Fundamental frequency and energy for a syllable were extracted every 10ms from a signal frame of 25ms multiplied with a Hamming window. Fundamental

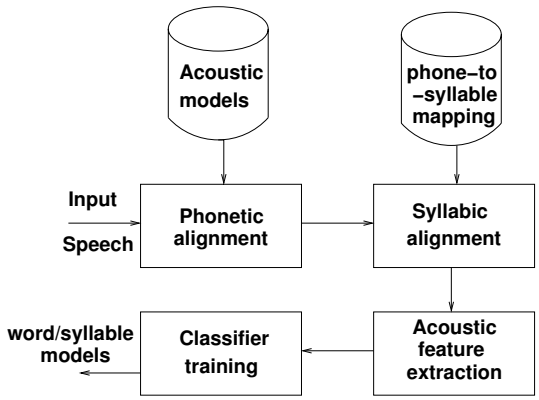

Fig. 1. Feature extraction and training of models

frequency of a frame was estimated using a high resolution pitch estimation algorithm described in [7]. Feature 4, *i.e.*, average filtered energy, is the average energy content in the higher frequency band of the spectrum which is used to incorporate a study done by Agaath *et. al.* in 1997 [8]. They concluded that the effect of stress in the energy content of the speech signal is more prominent in the higher band as compared to that in the lower band. Features 5 and 6 are derived from the basic prosodic features. These are obtained by multiplying average normalized energy and average normalized fundamental frequency of the syllable respectively with the normalized duration of the syllable. Feature 7 is the ratio of the average fundamental frequency of the next syllable to that of the current syllable. It reflects the change in the average fundamental frequency of two consecutive syllables. Similarly feature 8 is the ratio of average energy of the next syllable to that of the current syllable.

3. WORD DEPENDENT MODELS

Since the nature of lexical stress varies from one word to another, word dependent models can be built to determine if the word is spoken correctly from syllable stress viewpoint. In order to build a word dependent syllable stress model, acoustic features, extracted from all the constituent syllables of the word, are used to train the model. Since features 7 and 8 do not exist for the last syllable, the total number of features for a word is $(8 * N - 2)$ where N is the number of syllables in the word. All the acoustic features are concatenated to form a combined feature vector which is then used to train a particular classifier. We have built word dependent models for many different words and used four different classifiers, *viz.*, Naive Bayes, Decision Tree, k-NN and Support Vector Machine (SVM), to classify correctly and incorrectly spoken words.

4. WORD INDEPENDENT MODEL

We observe that it is possible to design standard classifiers such as decision tree or support vector machines to decide whether a word is spoken correctly or not. However, these standard classifiers are very specific to words. As a result, for each word we need one separate classifier. In practice, it has two major problems. First, with the increase in the number of words the number of required classifiers increase and therefore the model is not scalable at all for all practical purposes. Second, for a new word, it is always

Table 1. Syllabic level acoustic features

1.	Average fundamental frequency ($F0$)
2.	Average energy
3.	Duration
4.	Average filtered energy
5.	Average energy * duration
6.	$F0$*Duration
7.	$F0$ ratio
8.	Energy ratio

required to train a classifier from the very scratch. In other words, we cannot use the previously trained classifiers for new words. To circumvent these issues, we propose a novel technique mainly based on the naive Bayes classification paradigm.

Let us consider the two classes be ω and $\hat{\omega}$ representing the correct and incorrect classes such that if a word is spoken correctly then it belongs to ω, and it belongs to $\hat{\omega}$ otherwise. In order that a word is spoken correctly, each syllable of the word must follow the correct stress pattern in the word.

In order to have a generalized model, let there be a k-syllable word, $x : x_1x_2 \cdots x_k$ where x_i denotes the i^{th} syllable in the word. Let S denote the subset of syllables supposed to be stressed and US denote the subset of syllables to remain unstressed in x such that $S \bigcap US = \phi$ and $|US| = k - |S|$. In that case we can infer that the word is spoken correctly if and only if all syllables x_i, where $i \in S$, are stressed and all syllables x_j, where $j \in US$, are unstressed while speaking. Note that, in most of the English words $|S| = 1$ since we are considering only two levels of stress, *i.e.*, stressed and unstressed (secondary or tertiary stresses are treated as unstressed).

Now, let us consider two general classes of stressed and unstressed syllables independent of the words and denote them by C_s and C_{us} respectively. Note that, C_s and C_{us} are different from S and US in the sense that S and US denote the subsets of syllables to be stressed and unstressed in a particular word, x. In our formulation we establish a relation between the correct and incorrect word classes (ω and $\hat{\omega}$) and the general classes of stressed and unstressed syllables (C_s and C_{us}). In order for a word to belong to the correct class, ω, all syllables x_i where $i \in S$, should occur from the class C_s and for all syllables x_j where $j \in US$, should occur from the class C_{us}.

The posterior of a word belonging to the correct class can be denoted by

$$P(\omega|x) = \frac{P(x|\omega)P(\omega)}{P(x)} \tag{1}$$

Assuming independence of stress pattern for syllables, $P(\omega|x)$ can be written as

$$P(\omega|x) = \frac{\prod_{i=1}^{k} P(x_i|\omega)}{\prod_{i=1}^{k} P(x_i)} P(\omega) \tag{2}$$

In Equation 2, $P(x_i|\omega)$ represents the conditional probability that the syllable x_i belongs to the correctly spoken word, *i.e.*, the correctness of a word is governed by the correctness of each of its constituent syllables independently. A syllable will belong to correctly spoken word if the syllable is supposed to be stressed and it is actually stressed while speaking and vice versa. This is independent of other syllables and hence makes the independence

assumption valid. Let us consider that the class conditional probability models are available for C_s and C_{us}. In that case,

$$P(x_i|\omega) = \begin{cases} P(x_i|C_s) & \text{if } i \in S \\ P(x_i|C_{us}) & \text{if } i \in US \end{cases} \quad (3)$$

Therefore, the posterior of a word spoken correctly can be expressed as

$$P(\omega|x) = \frac{\prod_{i\in S}(P(x_i|C_s)\prod_{j\in US}P(x_j|C_{us}))}{\prod_{l=1}^{k}P(x_l)}P(\omega) \quad (4)$$

Since we can express $P(x_i) = P(x_i|C_s) + P(x_i|C_{us})$ (assuming that there are only two classes of stressed and unstressed syllables and no secondary or tertiary stressed syllables), the posterior can further be expressed as

$$P(\omega|x) = \prod_{i\in S}\left[\frac{P(x_i|C_s)}{P(x_i|C_s)+P(x_i|C_{us})}\right] \prod_{j\in US}\left[\frac{P(x_j|C_{us})}{P(x_j|C_s)+P(x_j|C_{us})}\right]P(\omega) \quad (5)$$

Let us denote

$$L(x_i|C_s) = \frac{P(x_i|C_s)}{P(x_i|C_s)+P(x_i|C_{us})} \quad (6)$$

and

$$L(x_i|C_{us}) = \frac{P(x_i|C_{us})}{P(x_i|C_s)+P(x_i|C_{us})} \quad (7)$$

as the likelihood ratios of a syllable x_i being generated from the class C_s and C_{us} respectively. We, therefore, can write

$$P(\omega|x) = \left(\prod_{i\in S}L(x_i|C_s)\prod_{j\in US}L(x_j|C_{us})\right)P(\omega) \quad (8)$$

Usually in naive Bayes formalism for two class classification

$$\begin{array}{ll} x \in \omega & \text{if } \ P(\omega|x) \geq 0.5 \\ x \in \hat{\omega} & \text{otherwise} \end{array} \quad (9)$$

However, since the size of the data is not very large, the *a priori* $P(\omega)$ can be a misleading one. Moreover, $P(\omega)$ is also word dependent. We, therefore, make decision by considering

$$\prod_{i\in S}L(x_i|C_s)\prod_{j\in US}L(x_j|C_{us}) \geq \theta \quad (10)$$

where θ is a threshold which depends only on the number of syllables in the word. For example, if x is a three syllable word with second syllable to be stressed, then x is spoken correctly, if $L(x_1|C_{us})L(x_2|C_s)L(x_3|C_{us}) \geq \theta$. This particular formulation makes the classification scheme capable of handling any new word provided the correct syllable stress pattern of the word is known. We have used GMM and C4.5 decision tree to model the class of stressed and unstressed syllables.

Table 2. Results for word dependent models. NB: Naive Bayes, DT: Decision Tree (C4.5), KNN: K Nearest Neighbours (K=3), SVM: Support Vector Machine

	Classification Accuracy (in percent)			
Word	NB	DT(C4.5)	KNN	SVM
AVAILABLE	94.28	92.85	94.28	87.5
CAPABILITY	65.83	70.73	78.04	68.29
CONDITION	87.83	75.75	86.48	75.75
CONTINUOUS	73.58	66.03	81.13	71.69
DETERMINE	94.91	84.74	91.52	93.22
DEVELOPED	84.21	80.70	89.47	73.68
EXPENSIVE	95.08	94.91	95.08	88.13
GOVERNMENT	71.15	59.61	53.84	55.76
INDUSTRY	84.12	88.88	88.88	80.95
INFORMATION	72.22	72.13	70.83	75.40
OPPOSITE	94.91	91.07	91.52	89.28
PROBABLY	76.27	74.57	81.35	79.66
REMEMBER	96.87	96.87	96.87	93.75
SUFFICIENT	64.40	64.40	59.32	64.40
TRADITIONAL	96.55	91.37	96.55	94.82
average word	84.07	78.31	84.14	77.83

5. EXPERIMENTS AND RESULTS

The experimental database consists of 25 words recorded in isolation by 53 speakers each at a sampling rate of 22 kHz in PCM format, resulting in 1325 utterances. All of the utterances were manually labeled by two human linguists, as correct or incorrect by considering the syllable stress used by the speaker. For feature extraction all the utterances were phonetically aligned using the acoustic models of an Indian English speech recognition engine built at IBM India Research Lab, New Delhi. The acoustic model is trained on more than 600 Indian speakers and can produce word accuracies of the order of 90% with speaker adaptation. We checked manually about 50% of the aligned utterances and corrections were made by hand.

Table 2 demonstrates the 10-fold cross-validation performance of four different word dependent classifiers namely naive Bayes, C4.5 decision tree, k-nearest neighbor and SVM. We used the implementation of WEKA [9] in studying the performance of all the classifiers. For k-NN classifier, we used $k = 3$ and for SVM, we used the sequential minimal optimization (SMO) [10] of WEKA with cubic polynomial kernel and $C = 1$ (C is the constant of SVM). As input to the word-dependent models, we used all the features of all the constituent syllables of a word to form a concatenated feature vector of dimension $(8 * N - 2)$. N is the number of syllables in the word which varies from 3 to 5 for different words in the training database.

For word-independent models, we used all the correctly spoken word utterances to collect instances of stressed and unstressed syllables across all the words. We assume that only the primary syllable is stressed in these utterances and rest of the syllables are unstressed. The word utterances which are labeled as incorrect can not be used for training as we do not have the information regarding which syllable (if any) is stressed in these utterances. A total of 2052 syllable utterances, containing 613 stressed and 1439 unstressed syllables, were used to train the GMMs and the C4.5 decision tree. In the case of GMM, each class is represented by a

Table 3. Results for word independent models

	Classification Accuracy (in percent)	
	GMM	Decision Tree
average syllable	82.23	92.16
Word		
AVAILABLE	90.00	94.28
CAPABILITY	65.85	73.17
CONDITION	85.13	75.67
CONTINUOUS	77.35	81.13
DETERMINE	61.01	89.83
DEVELOPED	61.40	64.91
EXPENSIVE	86.88	95.08
GOVERNMENT	65.38	73.07
INDUSTRY	85.71	88.88
INFORMATION	69.44	75
OPPOSITE	83.05	74.57
PROBABLY	71.18	77.96
REMEMBER	60.93	96.87
SUFFICIENT	66.10	83.05
TRADITIONAL	74.13	93.10
average word	72.12	80.25

mixture of 4 Gaussian components and full co-variance matrices. The trained decision tree consists of 51 nodes and 26 leaves. In the case of decision tree, we compute $L(x_i|C_s)$ for a test syllable x_i as $m_p/(m_p+n_p)$ where m_p and n_p are the number of stressed and unstressed syllables respectively assigned to p during training. p is the leaf node to which the test syllable x_i is assigned by the decision tree. To make the correct/incorrect decision for a word, the threshold θ was chosen as α^N, where N is the number of syllables in the word and α is a constant.

Table 3 demonstrates the results of word-independent models with $\alpha = 0.5$. We also varied α from 0 to 1 in steps of 0.05 in order to get the ROC curves as illustrated in Fig. 2. Average classification rates for isolated stressed and unstressed syllables are also shown in Table 3. We observe from Table 2 and Table 3 that the proposed word-independent model performs comparably with the word-dependent models. Table 3 also reveals that decision tree based word-independent model fares better than the GMM approach. This fact is also supported by the ROC curves as illustrated in Fig. 2. Overall Fig. 2 shows that we can indeed obtain a good region of operating characteristics using generic word-independent models instead of word dependent models.

6. CONCLUSION

In this paper, we proposed a generic word-independent model by establishing a relation between the class of stressed/unstressed syllables and the correct/incorrect spoken word utterances. Unlike the word dependent models, we can use the proposed word independent model to evaluate new words without requiring any additional training data. The experimental results show that the proposed word independent model is comparable with the word dependent models in terms of performance.

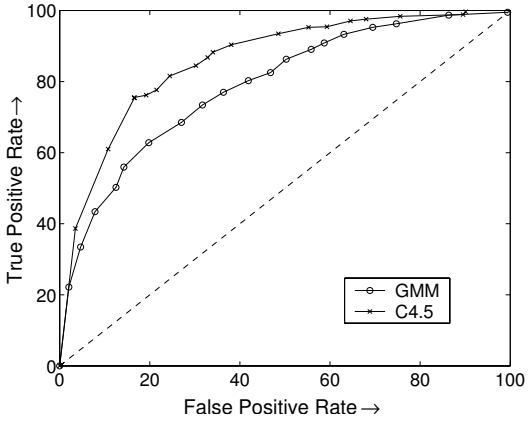

Fig. 2. ROC curves for word independent models for various threshold values

7. REFERENCES

[1] R. Silipo and Steven Greenberg, "Automatic transcription of prosodic stress for spontaneous english discourse," in *Proc. ICPh*, San Francisco, 1999, pp. 2351–2354.

[2] K. L. Jenkin and M. S. Scordilis, "Development and comparison of three syllable stress classifiers," in *Proc. ICSLP*, Philadelphia, 1996, pp. 733–736.

[3] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system," in *Proc. ICSLP*, Denver, Mar 2002, pp. 749–752.

[4] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. ICASSP*, Philadelphia, Mar 2005, pp. 733–736.

[5] R. Sarikaya and J. N. Gowdy, "Subband based classification of speech under stress," in *Proc. ICASSP*, May 1998, pp. 569–572.

[6] L. H. John Hansen and D. W. Brian, "Feature analysis and neural network-based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 307–313, Jul 1996.

[7] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 39, pp. 40–48, Jan 1991.

[8] M. C. S. Agaath, J. V. H. Vincent, and J. A. P. Jos, "Spectral balance as a cue in the perception of linguistic stress," *J. Acoust. Soc. America*, vol. 101, no. 1, pp. 503–513, Jan 1997.

[9] S. Garner, "Weka: The waikato environment for knowledge analysis," in *Proc. of the New Zealand Computer Science Research Students Conference*, 1995, pp. 57–64.

[10] J. C. Platt, "Sequential minimal optimization : A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, Microsoft Research, USA, 1998.