Model Based Stress Decision Method

Wooil Kim, Taeyun Kim, Sungjoo Ahn, Hanseok Ko
School of Electrical Engineering
Korea University, Korea
wikim@ispl.korea.ac.kr

Abstract
This paper proposes an effective decision method focused to evaluate the "stress position". Conventional methods usually extract the acoustic parameters and compare them to reference in absolute scale, adversely producing unstable results as testing condition changes. To cope with the environmental dependency, the proposed method is designed to be model-based that determines the stressed interval by making relative comparison over candidates. The stressed/unstressed models are then induced from normal phone models by adaptive training. The experimental results indicate that the proposed method is promising and that it is useful for automatic detection of stress positions. The results also show that generating the stressed/unstressed model by adaptive training is effective.

1. Introduction
As the growing surge in interest and research activities continue in speech recognition technology, many new applications are emerging. A case in point is the speech training systems for aiding students to effectively acquire foreign languages via speech pronunciation evaluation. Common automatic evaluation tools are known to employ a scheme similar to that used in speech recognition. For example, the quality of student's pronunciation is evaluated in terms of the probability score calculated from the trained acoustic models with the articulated speech signals applied. Such form of pronunciation training systems has been introduced as a service in internet and as computer games in CD media. However, all these systems introduced thus far lack the ability to evaluate the dominant acoustic features of pronunciation such as stress, intonation, and duration.

In an effort to analyze the features of articulation, various approaches have been introduced, among which is the method that extracts the acoustic parameters related to features and compares to reference pattern. This type of research effort has been pursued in laboratories primarily aimed at foreign language training and at developing tools in speech therapies [1].

In this paper, we propose a novel scheme to evaluate the stress among the pronunciation features to realize a fully automated speech training system. The conventional method for the detection of stress interval involves a procedure in which the acoustic features related to stress are extracted and compared to reference pattern. However, its performance is known to vary depending on testing environment. The proposed algorithm is based on speech model and uses a relative evaluation scheme, in order to produce more stable and robust results under varying acoustic environment.

The paper is organized as follows. We first identify the issues found in the conventional methods in Section 2. Section 3 reviews the Bayesian adaptive training technique used to generate the stressed/unstressed models. We then describe the proposed algorithm in detail in Section 4. The experimental procedures and results are presented and discussed in Section 5. Finally, in Section 6, we summarize the results and our findings in the form of concluding remarks.

2. Problem Formulation
In speech, stress is produced as a compounding effect of acoustic features such as pitch, intensity, and duration. It is observed that in stressed interval, pitch is increased, intensity is magnified and duration is elongated [2]. Consequently, for detecting the stressed interval, it is customary to extract these acoustic speech features and compare them to those of the reference pattern. However, we face difficulties in determining the stressed intervals using the normal feature extraction techniques because of the following reasons.

First, it is not easy to normalize the acoustic feature values. Feature values are not absolute and can vary erratically as the system condition changes. Thus, we need a scheme to normalize and transform the extracted feature values to render usable and valid measurements. Moreover, it is also difficult to estimate the acoustic variation. To make the matter worse, the variation is not consistent. In the case of pitch and formant, the vocal tract normalization (VTN) technique can be used as a possible speaker adaptation, but it requires the a priori information about speakers [3].

Second, it is also a formidable task to establish the absolute threshold used for decision rule. In common decision methods, the positions of stress interval are determined first by matching the extracted features with reference using selected pattern matching method such as dynamic time warping (DTW), and then by comparing the matching score against the threshold. In general, we can establish the threshold as a decision rule through the statistical analysis over a large amount of speech feature set. However, the feature values show unstable variations and the result of normalization is not concrete due to its varying tendency. As a result, it is unreliable to use the threshold found as absolute reference. These adversities prevent the ability in making objective decisions and consequently the system performance becomes unstable and depends heavily upon testing environment.

In addition, DTW automatically normalizes the time axis. Such property is disadvantageous to establishing reliable feature set since a change in duration as significant feature of stress cannot be properly reflected [4].

In this paper, we propose a model-based decision method in which the stress interval is adaptively determined relative to candidates. The stressed/unstressed vowel models are generated and to determine the stress interval, we make relative comparisons in the phone sequence candidates of the stress model combinations. To construct the stressed/unstressed model, a
database needs to be collected considering the acoustic occurrence of stress to each phones. However it is a hard work to make available sufficient speech data satisfying the requirement. In this paper, we also present a method to induce effective stressed/unstressed models from normal phone model with a small set of database.

3. BAYESIAN ADAPTIVE TRAINING

In speech recognition, adaptive training is used to improve the recognition performance by adjusting the acoustic model parameters to a new speaker or condition. In order to induce the stressed/unstressed models, we employ the MAP(Maximum A Posteriori) adaptive training method of the discrete HMM(Hidden Markov Model) parameters, in which the prior probability distribution of acoustic model is assumed. It is one of the Bayesian adaptive training techniques, where the prior probability density of the acoustic model is included in the estimation procedure.

If we assume that the parameter vectors \( \lambda = (\pi, A, B) \) of N-state discrete HMM are independent and the rows of them are independent, their probability densities have Dirichlet distributions while the prior probability density function \( g(\lambda) \) shows a special case of the matrix beta probability density function as follows [5].

\[
g(\lambda) = K_\pi \prod_{i=1}^{N} \left( \pi_i^{\eta_i - 1} \cdot \prod_{j=1}^{N} a_{ij}^{\nu_{ij} - 1} \cdot \left( \prod_{k=1}^{K} \beta_{ik}^{\nu_{ik} - 1} \right) \right) \quad (1)
\]

where \( K_\pi \) is the constant for normalization and \( \eta_i, \nu_{ij} \), and \( \nu_{ik} \) are the hyper parameters representing the probability density function.

Given the observation sequence \( x \) and the prior probability density \( g(\lambda) \), the MAP estimation including the prior probability density is

\[
\lambda_{MAP} = \arg \max_{\lambda} P(x|\lambda)g(\lambda) \quad (2)
\]

Applying the EM(Expectation Maximization) algorithm to Equation (2), each parameters can be estimated as follows:

\[
\hat{\pi}_i = \frac{\epsilon_i + \eta_i - 1}{\sum_{i=1}^{N} (\epsilon_i + \eta_i - 1)} \quad i = 1, 2, \ldots, N \quad (3)
\]

\[
\hat{a}_{ij} = \frac{c_{ij} + \eta_j - 1}{\sum_{j=1}^{N} (c_{ij} + \eta_j - 1)} \quad i, j = 1, 2, \ldots, N \quad (4)
\]

\[
\hat{\beta}_{jk} = \frac{d_{jk} + \nu_{jk} - 1}{\sum_{k=1}^{K} (d_{jk} + \nu_{jk} - 1)} \quad j = 1, 2, \ldots, N, k = 1, 2, \ldots, K \quad (5)
\]

where \( \epsilon_i, c_{ij} \) and \( d_{jk} \) are defined as:

\[
\epsilon_i = P(r_i = 1|x, \lambda) \quad (6)
\]

\[
c_{ij} = \sum_{t=1}^{T} P(r_{t+i} = 1, r_{t+i+1} = j|x, \lambda) \quad (7)
\]

\[
d_{jk} = \sum_{t:x_t \sim \nu_k} P(r_{t+j} = 1, x_t \sim \nu_k|x, \lambda) \quad (8)
\]

\( \eta_i \) is the state at time \( t \) and \( x_t \sim \nu_k \) means the observation vector \( x_t \) is corresponding to the output symbol \( \nu_k \).

To estimate the HMM parameters from Equation (2), it needs to find the hyper parameters \( \eta_i, \eta_j, \nu_{jk} \), which are related with the prior probability density function \( g(\lambda) \). These hyper parameters can be found using two possible approaches. The first approach is to estimate the parameters from the extracted statistical information such as mean and variance from the available acoustic models under exhaustive acoustic conditions as in various speakers or environment. It can be well used when we have the HMM models of each speaker or environment. The second approach is to find the hyper parameters from merged database with each individual speaker or condition. The estimated counts \( \hat{\epsilon}_i, \hat{c}_{ij}, \hat{d}_{jk} \) can be obtained by the HMM parameters estimated from the merged database such as speaker independent model. They are divided by training tokens of each speech unit and are added by one and then used for establishing new hyper parameters \( \eta_i, \eta_j, \nu_{jk} \) [5].

In this paper, employing the latter one, we find the hyper parameters from the normal phone models and induce the stressed/unstressed vowel models from small amount of the stress-considered database using the hyper parameters.

4. STRESS DECISION BASED ON STRESS/UNSTRESSED MODEL

First, the phone models are estimated from the normal speech database in which the stress events are not considered. Then the stress reflected acoustic models by applying the Bayesian adaptive technique described in Section 3 to the database constructed by considering the stress occurrence. The database also contains the dictionary with stress information labeled. The hyper parameters, essential to adaptive training, are extracted from the normal phone models. We now have both stressed and unstressed models to every vowel. Figure 1 shows the procedure of generation of stressed/unstressed models.

Since most speech training system simply prompts user to utter a certain word/sentence, it is known to the system what pronunciation is anticipated as response. After receiving the spoken word as input, the basic phone sequence corresponding to the input speech is composed. As a result, the candidate phone sequences are generated for a duration as long as the total syllables of the spoken word with the stressed vowel model in one syllable and the unstressed in the other. With the HMM models matching to the candidate sequences, the probability scores are computed using the Viterbi decoding. Among the scores computed, the stress interval is decided by the position of the stress model in the candidate corresponding to the maximum score. (Figure 2)
5. EXPERIMENTAL RESULTS

We ran several representative experiments using PBW452 database, which is a Korean speech corpus of the Korean Information Base System. As a ground work, we constructed the isolated word recognition system based on the speaker independent phone model. The baseline system has the specification and performance as delineated in Table 1.

First, we conducted a speaker adaptation experiment to verify the performance of the adaptive training method to be used for stress model generation. For performance comparison, we trained 4 types of the acoustic model as follows.

1. SI1: speaker independent model (4520 words of 10 speakers)
2. SI2: speaker independent model trained including the test speaker data (4520 words of 10 speakers and 452 words of the test speaker)
3. SD: speaker dependent model trained from SI1 as the initial model (452 words of the test speaker)
4. SA: speaker adapted model induced by Bayesian adaptive training from SI1 (452 words of the test speaker)

As shown in Table 2, the adapted model shows the best performance in the recognition experiment. It is generally known that speaker dependent system is superior over independent system. However, in case of SD, the insufficient database (only 452 samples) makes the SD model sharply degrade compared to speaker independent models (SI1, SI2). From the experiment, we could ascertain that the employed adaptation technique improves the recognition performance even in an insufficient database by adapting the model to specific condition. These results indicate that it is reasonable to generate the stressed/unstressed model from normal phone models using adaptive technique with small amount of stress-reflected speech data.

To train the stressed/unstressed model, we need a speech database in which the stress events are considered. We constructed a Korean stress-considered speech database in the following way. We constructed a word set which consists of 21 words containing 7 principal vowels (/a/, /æ/, /i/, /ɪ/, /u/, /ʊ/, /ʌ/) equally and collected 504 words $8 \times 3 \times 21$ by having eight male speakers articulate every word and placing stress at each syllable. Among them, 378 words by six speakers are used for adaptive training and 98 words including principal vowels of other two speakers are used for stress decision test.

The stressed/unstressed models of principal vowels were obtained from adaptive training. The decision on stress position is made in accordance to the following procedure. First, three candidate phone sequences having stressed vowel model at each syllable respectively are constructed to establish the words to be tested. For example, when deciding the stress position of a word spoken as in /duun/ (which means “dark” in English), we construct the first candidate phone sequence in which the first syllable /d/ reflects a stressed model while the other two syllables /u/ and /un/ reflect unstressed models. Also, we create the second candidate using only with the stressed model at the second syllable /d/ and the last one with the model at third syllable /un/. We then computed the maximum probability scores of candidate with input speech using the Viterbi search. Finally, we made decision on the final evaluation according to the stress position of the candidate with the highest score. The results of the stress decision experiment are shown in Table 4.

We obtained 76.53% as the success rate from the proposed scheme on the decision performance of stress position. Although the experimental results show a possibility as useful means for automatic stress decision, it falls short of our expectation. We can reason out the low performance result as follows. First, the stress is one of the supra-segmental features. The acoustic events of the stress occur not in vowels alone but include other adjacent phones. One syllable interval, initially designed to reflect the stressed model in experiment, could not exhibit the supra-segmental property effectively. Second, we found that some syllables with aspirations or fortises as adjacent consonants are articulated forcefully and consequently they become obscure under the other stressed syllables. These phenomena have to be considered for further performance improvement in future work. Finally, in Korean language, the stress is not a noticeable feature as other tone languages such as English or Chinese, especially in isolated word articulations. In addition, the stress speech data used in the experiments were pronounced artificially, so that some spoken words may have failed to show natural stress events. For example, we observed that some syllables were articulated only with high intensity maintaining the pitch
uniformly or only with the duration elongated. Such speech data show adverse effects to adaptive training, preventing adequate model generations.

From the experiments presented above, we can ascertain that it is reasonable to determine the stress interval using stressed/unstressed model and to induce the model from insufficient database by adaptive training. Since the proposed algorithm is designed to adaptively evaluate the spoken word based on the probability scores of candidate phone sequences, it can compensate the conventional method, wherein an absolute comparison is made to the reference and shows unstable performance as the testing condition changes.

6. CONCLUSIONS

In this paper, we have proposed an effective decision method to evaluate the stress position to cope with the unstable performance shown in conventional method. The proposed method is an adaptive evaluation scheme reflecting the model condition changes. The experimental results indicate that the proposed method shows promising results useful for automatic detection of stress positions.

As a future work, we plan to improve the method by reflecting the supra-segmental properties incorporating stressed/unstressed models to other phones as well as principal vowels.

7. References