INTERSPEECH 2009 BRIGHTON

# Word stress assessment for computer aided language learning

*Juan Pablo Arias*[1], *Nestor Becerra Yoma*[1], *Hiram Vivanco*[2]

[1]Speech Processing and Transmission Laboratory
Department of Electrical Engineering, Universidad de Chile, Santiago, Chile
[2]Department of Linguistics, Universidad de Chile, Santiago, Chile
juaarias@ing.uchile.cl, nbecerra@ing.uchile.cl

## Abstract

In this paper an automatic word stress assessment system is proposed based on a top-to-bottom scheme. The method presented is text and language independent. The utterance pronounced by the student is directly compared with a reference one. The trend similarity of $F0$ and energy contours are compared frame-by-frame by using DTW alignment. The stress assessment evaluation system gives an EER equal to 21.5%, which in turn is similar to the error observed in phonetic quality evaluation schemes. These results suggest that the proposed system can be employed in real applications and applicable to any language.

**Index Terms**: word stress assessment, computer-aided language learning

## 1. Introduction

Undoubtedly, speech technology has played an important role in computer-aided language learning (CALL). The traditional paradigms like laboratory audio tapes have been replaced by more natural interactions. The old systems based on static pictures are replaced by real dialogues where it is possible to evaluate pronunciation quality or fluency. As a consequence, CALL systems provides several advantages to students and the learning process takes place in a more motivating context characterized by interactivity, motivation and even entertainment. Also, students usually feel inhibited about speaking out in class and CALL can provide a more convenient environment to practise a second language.

Despite the fact that most of students of English as a second language may achieve acceptable writting and reading skills, their pronunciation show poor quality, fluency and naturalness. Certainty, the phonetic rules take most of the attention in the learning process of oral communication skills. However, prosody is probably the most important aspect to achieve a natural and fluent pronunciation when compared with native speakers. Among the prosodic or suprasegmental features, intonation and stress are probably the most relevant.

Word stress depends on the intensity with which a sound is pronounced. The presence of syllables receiving a main or a secondary stress is important in English as the segments in them tend to be pronounced fully. Weakening and vowel reduction usually occur in unstressed syllables, phenomena that are not so marked in Spanish. Stress placing may change the meaning of a word. For example, the word "object" is a noun if stressed on the first syllable and a verb if stressed on the second. There are interesting cases in word compounds in English in which stress is significant: pairs like "the white house" (a house which is white) and "the Whitehouse" (the residence of the Presidents of the USA), for instance. Therefore, a stress mistake may generate confusion or misunderstanding and obstruct the communication.

Surprisingly, the problem of word stress assessment from second language learning point of view has not been addressed exhaustively in the literature. Most of the papers on pronunciation quality assessment have addressed the problem of phonetic quality evaluation [1, 2, 3]. Some authors have used prosodic features like intonation as an additional variable to assess pronunciation quality in combination with other features [4, 5]. A prosodic module (including stress activities) for foreign language learning is presented in [6]. Moreover, the system requires human assistance to insert orthographic information. In [7], an automatic syllable stress detection system is presented. However, the classification is performed syllable-by-syllable and hence the text transcription of the reference utterance is required.

The proposed system is not text-dependent and minimizes the effect of the resulted phonetic quality in the student's utterance. The word stress evaluation system that results from the combination of $F0$ and energy contour estimation provides an EER equal to 21.5%, which in turn is comparable to the error of phonetic quality pronunciation assessment systems. Despite the fact that the system introduced here was tested with the English language, it can be considered as language-independent.

## 2. The proposed system

The system attempts to decide, on a top-to-bottom basis, if two utterances (i.e. reference and testing ones), from different speakers, were produced with the same stress pattern. Figure 1 [8] shows the block diagram of the proposed scheme to assess the stress generated by a student of a second language. First, prosodic features (frame energy and $F0$), and Mel-frequency cepstral coefficients (MFCC) are estimated in both utterances. The $F0$ contours are represented in the log domain, normalized with respect to the mean value and smoothed to allow the comparison of $F0$ curves from different speakers (e.g. a male and a female). Also, the frame energy contour is estimated and represented in the log scale. Then both sequences of MFCC parameters are aligned by using a standard DTW alignment. Finally, the reference and testing $F0$ and frame energy curves are compared on a frame-by-frame basis by employing the DTW alignment obtained with the MFCC observation sequences.

### 2.1. $F0$ contour extraction and post-processing

First, the speech signals are sampled at 16 kHz. An end-point detection and a high-pass filter at 75 Hz cutoff frequency are applied. Then, a pre-emphasis is applied. After, speech signals are low pass filtered at 600Hz cutoff frequency and divided

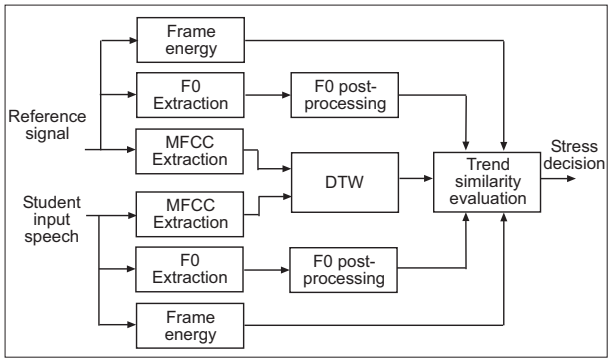


Figure 1: *Block diagram of the proposed system.*

into 400-sample frames with a 50% overlapping. Then, $F0$ is estimated at each frame and represented in a semitone scale according to:

$$F0_{semitone}(t) = 12\frac{\log F0(t)}{\log 2} \tag{1}$$

where $F0(t)$ and $F0_{semitone}(t)$ are, respectively, the fundamental frequency in Hertz and in the semitone scale adopted here at frame $t$. Curve $F0_{semitone}(t)$ is smoothed with a median filter. Then it is normalized with respect to the mean value. Finally, the discontinuities caused by unvoiced intervals are filled by linear interpolation. The resulted post-processed $F0$ curve is denoted by $F0_{pp}(t)$.

### 2.2. Energy contour extraction

The energy (intensity) contour extraction is included and combined with the post-processed $F0$ curve to decide if the stress in the reference utterance is the same as the testing one. The energy contour at frame $t$, $E(t)$, is estimated as:

$$E(t) = 10 \cdot log\left[\sum_{n=1}^{N} x^2(t+n)\right] \tag{2}$$

where $x(\cdot)$ denotes the signal samples and $N$ is the frame width.

### 2.3. DTW based alignment

Thirty-three MFCC parameters per frame are computed in the reference and testing utterance: the frame energy plus ten static coefficients and their first and second time derivatives. Then, DTW algorithm is applied to align both observation sequences. Local distance between frames is estimated with Mahalanobis metric. The resulted optimal alignment provided by DTW is indicated by $I(k) = \{i_R(k), i_S(k)\}, 1 \leq k \leq K$ where $i_R(k)$ and $i_S(k)$ are the index of frames from the reference and testing utterance, respectively, which are aligned. As it is well known in the literature, the accuracy of DTW-based speech recognition systems is dramatically degraded when the speaker matching condition is not valid. However, the proposed method in this paper employs the DTW-based alignment instead of the DTW-based global metrics as in speech recognition systems. As shown here, speaker mismatch condition, which can also result in a quality pronunciation mismatch, has a restricted effect in the optimal alignment and in the overall system accuracy.

### 2.4. Stress similarity assessment

According to Fig. 1, the trend similarity between the reference and testing post-processed $F0$ and energy contour is estimated. As described above, the comparison is done on a frame-by-frame basis using DTW alignment. However, instead of just estimating the accumulated distance between reference and testing utterances, this paper proposes that the prosody of both utterances should be compared from the falling-rising trend point of view. In other words, the system should decide if the student is able to produce a $F0$ and energy contours with the same falling-rising pattern as the reference utterance. If $[E^R(t), F0^R_{pp}(t)]$ and $[E^S(t), F0^S_{pp}(t)]$ denote the pairs energy contours and post-processed $F0$ curves from reference and student's testing utterances, respectively, the trend similarity measure $TS(F0^R_{pp}, E^R, F0^S_{pp}, E^S)$ is computed as:

$$\begin{aligned} &TS(F0^R_{pp}, E^R, F0^S_{pp}, E^S) = \\ &\alpha TS(E^R, E^S) + (1-\alpha)TS(F0^R_{pp}, F0^S_{pp}) \end{aligned} \tag{3}$$

where $TS(E^R, E^S)$ and $TS(F0^R_{pp}, F0^S_{pp})$ are estimated by making use of the correlation of $E^R$ with $E^S$, and of $F0^R_{pp}$ with $F0^S_{pp}$, respectively. Given the DTW alignment between the reference and testing utterances, $I(k)$, $TS(E^R, E^S)$ and $TS(F0^R_{pp}, F0^S_{pp})$ are computed as [8]:

$$TS[E^R(t), E^S(t)] = \frac{\sum_{k=1}^{T}\left\{E^R[i_R(k)] - \overline{E^R}\right\}\left\{E^S[i_S(k)] - \overline{E^S}\right\}}{\sigma_{E^R} \cdot \sigma_{E^S}} \tag{4}$$

$$TS[F0^R_{pp}(t), F0^S_{pp}(t)] = \frac{\sum_{k=1}^{T}\left\{F0^R_{pp}[i_R(k)] - \overline{F0^R_{pp}}\right\}\left\{F0^S_{pp}[i_S(k)] - \overline{F0^S_{pp}}\right\}}{\sigma_{F0^R_{pp}} \cdot \sigma_{F0^S_{pp}}} \tag{5}$$

where $\sigma_{E^R}$, $\sigma_{F0^R_{pp}}$, $\sigma_{E^S}$ and $\sigma_{F0^S_{pp}}$ are the standard deviation of $E^R$, $F0^R_{pp}$, $E^S$ and $F0^S_{pp}$, respectively. Finally, the system takes the decision about the stress pattern resulted from the student's utterance, $SD$, according to:

$$SD\left[TS(F0^R_{pp}, E^R, F0^S_{pp}, E^S)\right] = \begin{cases} \text{same stress} & \text{if } TS(F0^R_{pp}, E^R, F0^S_{pp}, E^S) \geq \theta_{SD} \\ \text{different stress} & \text{elsewhere} \end{cases} \tag{6}$$

where $\theta_{SD}$ corresponds to a decision threshold.

## 3. Experiments

### 3.1. Database

A database was recorded at the Speech Processing and Transmission Laboratory (LPTV), Universidad de Chile, to evaluate the performance of the proposed scheme to address the problem of stress. All the speech material was recorded in an office environment with a sampling frequency equal to 16 kHz. There are two types of speakers: the experts and the non-experts in English language and phonetics. The expert speakers correspond to a professor of English language and his last-year students at

the Department of Linguistics at Universidad de Chile. All the non-expert speakers demonstrated an intermediate proficiency in English. Three microphones were employed: Shure PG58 Vocal microphone (Mic1) and two low-cost desktop PC microphones (Mic2 and Mic3).

This data set is composed by twelve words: "machine"; "alone"; "under"; "husband"; "yesterday"; "innocence"; "important"; "excessive"; "melancholy"; "caterpillar"; "impossible"; and, "affirmative". Each word was uttered with all the possible stress variants, which in turn are word dependent. The average number of stress variants is equal to three patterns per word. Altogether there are 12 sentences $\times$ 3 stress patterns = 36 types of utterances that were recorded by eight speakers (four experts and four non-experts in English language and phonetics) by making use of three microphones simultaneously. Then, the total number of recorded sentences is equal to 36 types of utterances $\times$ 8 speakers $\times$ 3 microphones = 864 utterances. In the stress assessment experiment, the reference utterances correspond to sentences recorded by one of the experts in English language and phonetics (the most senior one). Finally, the total number of stress assessment experiments is equal to 36 experiments per speaker per microphone $\times$ 7 testing speakers $\times$ 3 microphones = 756 experiments.

### 3.2. Experimental setup

The DTW algorithm mentioned in Fig. 1 was implemented according to [9]. The fundamental frequency $F0$ is estimated by using the autocorrelation based Praat pitch detector system [10]. As mentioned above, the utterances are divided into 400-sample frames with a 50% overlapping. Thirty-three MFCC parameters per frame were computed: the frame energy plus ten static coefficients and their first and second time derivatives.

### 3.3. DTW alignment accuracy experiments

The speaker mismatch effect on DTW accuracy alignment is evaluated in this paper. A subset of three expert speakers and two non-expert speakers were selected to record 24 sentences, to assess the robustness to speaker and pronunciation quality mismatch of the DTW alignment. The utterances recorded with two microphones were employed: Shure PG58 Vocal microphone and one of the low-cost desktop PC microphones. Therefore, a total number equal to 240 utterances were used. These utterances were phonetically segmented and labelled by hand. The alignment error at phonetic label border $b$, $E_{align}(b)(\%)$, is defined as:

$$E_{align}(b) = 100 \cdot \frac{d(b)}{D} \tag{7}$$

where $D$ is the searching windows width in DTW, and $d$ is defined as:

$$d(b) = \frac{1}{2}\sqrt{d_R^2(b) + d_S^2(b)} \tag{8}$$

where $d_R(b)$ and $d_S(b)$ are the horizontal and vertical distances, respectively, between the phonetic boundaries obtained by hand-labelling and the DTW alignment (See Fig. 2). Given two utterances with the same text transcription, the total alignment error, $E_{align}$ , is equal to:

$$E_{align} = \frac{1}{B}\sum_{b=1}^{B} E_{align}(b) \tag{9}$$

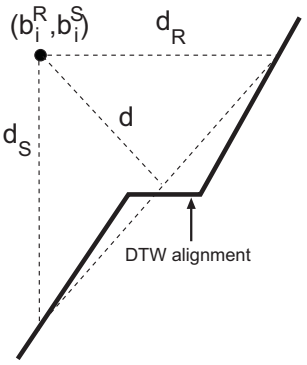


Figure 2: *Representation of DTW alignment error measure, d. Point $(b_i^R, b_i^S)$ indicates the intersection of boundary i within the reference and testing utterances. The distances $d_R$ and $d_S$ are the horizontal and vertical distances, respectively, between the phonetic boundaries and the DTW alignment.*

where B is the total number of phonetic boundaries in the sentences.

## 4. Results and Discussion

Table 1 compares the DTW alignment error between speaker matched and unmatched condition with data set described in subsection 3.3. As can be seen, when compared with speaker matching condition, the alignment error shows an increase of just 1.36 percentage points when reference and testing utterances are recorder by different speakers. Table 2 shows the alignment error achieved with pronunciation quality matching and mismatching conditions between the reference and testing utterances. As can be seen, when compared with pronunciation quality matching condition, the alignment error shows an increase of just 0.62 percentage points when phonetic quality in the student's testing utterance is reduced. Consequently, despite the fact that the DTW-based speech recognizer system accuracy dramatically degrades with mismatch condition between reference and testing utterances, results in Tables 1 and 2 strongly suggest that the DTW alignment is robust to speaker and pronunciation quality mismatch.

Table 1: *Alignment error with speaker matched and unmatched condition.*

| **Speaker matching condition** | **Alignment error** |
|---|---|
| *Matched* | 2.86% |
| *Unmatched* | 4.22% |

Table 2: *Alignment error with pronunciation quality matched and unmatched condition.*

| **Pronunciation quality condition** | **Alignment error** |
|---|---|
| *Matched* | 4.10% |
| *Unmatched* | 4.72% |

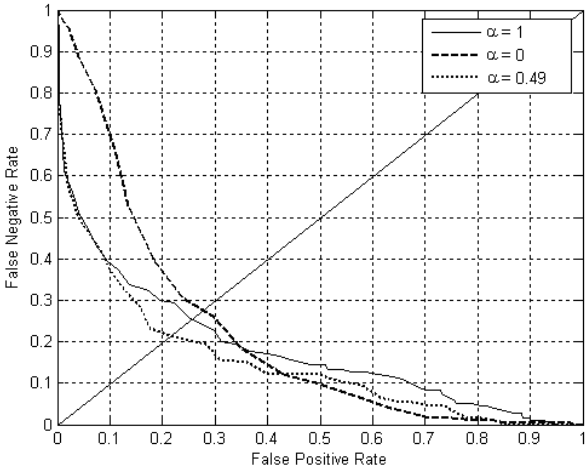


Figure 3: *False negative and false positive ROC curves in stress evaluation. The trend similarity measure is estimated according to (4) and the decision is taken by using (5). $\alpha = 1$ indicates that only pitch contour is employed and $\alpha = 0$ indicates that only frame energy contour is employed.*

Figure 3 presents the receiver operating characteristic (ROC) curves (false negative rate, FNR, and false positive rate, FPR) with the stress assessment system shown in Fig. 1. The trend similarity is estimated with (6) and the final decision about stress assessment is taken according to (6). The variable $\alpha$ was tuned in order to minimize the area below the ROC curve and the optimal value is equal to 0.49. Figure 3 also shows the FPR/FNR curves with $\alpha = 0$; $\alpha = 1$; and, $\alpha = 0.49$. Table 3 presents the area below the ROC curve and EER with $\alpha$ equal to 0, 1 and 0.49. According to Fig. 3 and Table 3, the optimal $\alpha$ gives a reduction in the area below the ROC curve and in EER equal to 15.5% and 22.3%, respectively, when compared with $\alpha = 0$ and $\alpha = 1$. This result suggests that both pitch and energy contours provide relevant information to assess word stress. The stress assessment system accuracy should be improved by including duration information, which in turn is not straightforward in the frame of the DTW alignment. However, it is worth highlighting that the optimal EER equal to 21.5% is similar to phonetic pronunciation assessment system that suggests that the proposed scheme is accurate enough for practical applications.

Table 3: *ROC area and equal error rate (EER) for stress assessment system for different $\alpha$, using correlation as trend similarity measure.*

| Feature | ROC area | EER (%) |
|---|---|---|
| $\alpha = 1$ | 0.181 | 25.4 |
| $\alpha = 0$ | 0.212 | 27.6 |
| $\alpha = 0.49$ | 0.147 | 21.5 |

# 5. Conclusions

In this paper a word stress assessment system based on a top-to-bottom scheme is presented. The system compares the utterance pronounced by the student with a reference one. The trend similarity of $F0$ and energy contours are compared on a frame-by-frame basis by using the DTW alignment. Also, the robustness of the alignment provided by the DTW algorithm to pronunciation quality and speaker mismatch is addressed. The stress assessment evaluation system provides an EER equal to 21.5%, which in turn is similar to the error observed in phonetic quality evaluation schemes. These results suggest that the proposed systems can be employed in real applications. Despite the fact that the system was tested in the framework of English learning with native-Spanish learners, the proposed method is applicable to any language. Finally, the use of techniques to improve robustness to noise, and the integration of the schemes proposed in this paper with phonetic quality and duration evaluation are proposed as future research.

# 6. Acknowledgements

This work was funded by Conicyt-Chile under grants Fondef No. D05I-10243 and Fondecyt No. 1070382.

# 7. References

[1] Neumeyer, L., Franco, H., Weintraub, M., and Price, P., (1996) "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech", in Proc. ICSLP '96.

[2] Franco, H., Neumeyer, L., Kim, Y. and Ronen, O., (1997) "Automatic pronunciation scoring for language instruction", in ICASSP'97, 1997, vol. 2, pp. 1471-1474.

[3] Gu, L., Harris, J., (2003) "SLAP: a system for the detection and correction of pronunciation for second language acquisition" in International Symposium on Circuits and Systems, ISCAS '03, Vol. 2, pp. 580-583.

[4] Dong, B., Zhao, Q., Zhang, J., Yan, Y., (2004) "Automatic assessment of pronunciation quality" in International Symposium on Chinese Spoken Language Processing, Dec. 2004, pp. 137-140.

[5] Weiqian, L., Jia, L., Runsheng, L., (2005) "Automatic spoken English test for Chinese learners" in Proceedings of International Conference on Communications, Circuits and Systems, 2005,Volume: 2, pp. 860-863.

[6] Delmonte, R., Peterea, M., Bacalu, C., (1997) "SLIM: Prosodic Module for Learning Activities in a Foreign Language", in Proc. ESCA, Eurospeech 97, Rhodes, Vol. 2, pp. 669-672.

[7] Tepperman, J., Narayanan, S., (2004) "Automatic Syllable Stress Detection Using Prosodic Features for Pronunciation Evaluation of Language Learners", in ICASSP-2005, vol. 1, pp. 937-940.

[8] Arias, J.P., Yoma, N.B., Vivanco, H., (2009) "Automatic Intonation assessment for Computer-aided Language Learning", submitted to Speech Communication (Elsevier), January, 2009.

[9] Sakoe, H. and Chiba, S., (1978) "Dynamic programming algorithm optimization for spoken word recognition" in IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26.

[10] Boersma, P., Weenink, D., (2008) Praat: doing phonetics by computer (Version 5.0.29) [Computer program]. Retrieved July 14, 2008, from http://www.praat.org/