Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 254–267
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Automatic intonation assessment for computer aided language learning

Juan Pablo Arias [a], Nestor Becerra Yoma [a,*], Hiram Vivanco [b]

[a] Speech Processing and Transmission Laboratory, Department of Electrical Engineering, Universidad de Chile, Av. Tupper 2007, P.O. Box 412-3, Santiago, Chile
[b] Department of Linguistics, Universidad de Chile, Santiago, Chile

Received 29 January 2009; received in revised form 10 November 2009; accepted 12 November 2009

Abstract

In this paper the nature and relevance of the information provided by intonation is discussed in the framework of second language learning. As a consequence, an automatic intonation assessment system for second language learning is proposed based on a top-down scheme. A stress assessment system is also presented by combining intonation and energy contour estimation. The utterance pronounced by the student is directly compared with a reference one. The trend similarity of intonation and energy contours are compared frame-by-frame by using DTW alignment. Moreover the robustness of the alignment provided by the DTW algorithm to microphone, speaker and quality pronunciation mismatch is addressed. The intonation assessment system gives an averaged subjective–objective score correlation as high as 0.88. 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 systems could be employed in real applications. Finally, the schemes presented here are text- and language-independent due to the fact that the reference utterance text-transcription and language are not required.

Keywords: Intonation assessment; Computer aided language learning; Word stress assessment

1. Introduction

Computer aided language learning (CALL) has replaced the traditional paradigms (e.g. laboratory audio tapes) with human–machine interfaces that can provide more natural interactions. The old systems based on static pictures are replaced by real dialogues between the user and the system, where it is possible to evaluate pronunciation or fluency quality and to input answers by voice. In this new paradigm, speech technology has played an important role. As a result, CALL systems provide several advantages to students and the learning process takes place in a more motivating context characterized by interactivity (Traynor, 2003). Moreover, students usually feel inhibited about speaking out in class (Bernat, 2006) and CALL can provide a more convenient environment to practise a second language.

The suprasegmental characteristics of speech (pitch, loudness and speed) (Wells, 2006) are very important issues when learning a foreign language. For instance, most students of English as a second language may achieve acceptable writing and reading skills, but their pronunciation may not reach the same standard, lacking fluency and naturalness, among other characteristics. It is worth mentioning that for some authors naturalness of style implies fluency. For instance, according to (Moyer, 2004), "The extent of contextual isolation, or even text type itself, may evoke varying degrees of naturalness in style, and therefore fluency." Moreover, sometimes teachers show poor oral skills themselves (Gu and Harris, 2003; Baetens, 1982), which in turn is an additional barrier to beginner students. Despite the fact that the phonetics rules (understood as rules for the correct pronunciation of segments (Saussure et al., 2006; Holmes and Holmes, 2001; El-Imam and Don, 2005)) take most of the attention in the learning process of oral communication skills, in the case of more

* Corresponding author.
E-mail address: nbecerra@ing.uchile.cl (N.B. Yoma).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.11.001

advanced students, producing the correct prosody is probably the most important aspect (Delmonte et al., 1997) to achieve a natural and fluent pronunciation when compared with native speakers. In this context, speech analysis plays an important role to help students to practise and improve their oral communication skills, without the need of teacher assistance (Rypa and Price, 1999). Also, providing adequate feedback is a very relevant issue in CALL (Chun, 2002) because it can motivate students to practise and improve their pronunciation skills. Furthermore, there is strong evidence that audiovisual feedback improves the efficiency of intonation training (Botinis et al., 2001; Shimizu and Taniguchi, 2005).

In the context of prosody, intonation is certainly targeted more often than energy and duration in second language learning. Intonation is strongly related to naturalness, emotional colour and even meaning as it is explained later in Section 2. Also, word accent results in *F*0 movements which play a role in the syllable stress mechanism (You et al., 2004). The problem of intonation assessment has been addressed from several points of view: nativeness assessment; fluency evaluation and training; classification; and, computer aided pronunciation quality evaluation. In (Tepperman and Narayanan, 2008; Teixeira et al., 2000) text-independent based methods are employed to evaluate the degree of nativeness by analyzing the pitch contour. In (Eskenazi and Hansma, 1998) a fluency pronunciation trainer strategy is presented by assessing prosodic features. Firstly, the user is prompted to repeat a given sentence and duration is corrected separately from the other features, then the user can proceed to pitch, etc. The duration information is provided by forced Viterbi alignment (Jurafsky and Martin, 2009). It is worth highlighting that the forced Viterbi algorithm can automatically estimate the phoneme boundaries given the utterance text transcription. In (Peabody and Seneff, 2006) an automatic tone correction in non-native Mandarin is proposed by comparing user-independent native tone models with the pitch contours generated by the students. Observe that in (Tepperman and Narayanan, 2008; Teixeira et al., 2000; Eskenazi and Hansma, 1998; Peabody and Seneff, 2006) a bottom-up philosophy is employed to evaluate prosodic features by using text-independent or user-independent models. Moreover, observe that the intonation assessment problem from the CALL point of view is not necessarily a nativeness evaluation, fluency evaluation nor pitch contour classification problem with predefined classes.

Surprisingly, the problem of pronunciation quality evaluation from the intonation point of view in second language learning has not been addressed exhaustively in the literature. Most of the papers on pronunciation quality assessment have addressed the problem of phonetic quality evaluation (Neumeyer et al., 1996; Franco et al., 1997; Gu and Harris, 2003). However, some authors have used intonation as an additional variable to assess pronunciation quality in combination with other features (Dong et al., 2004; Liang et al., 2005). In (Delmonte et al., 1997), a prosodic module (including intonation and stress activities) for foreign language learning is presented. The system compares the student's utterance with a reference one by using a heuristic based approach. Moreover, the system requires human assistance to insert orthographic information and does not provide any scoring. In (Kim and Sung, 2002; You et al., 2004) an intonation quality assessment system is proposed based on a bottom-up scheme where the intonation curve within each syllable is classified. The system makes use of forced Viterbi alignment and hence is text-dependent. In (van Santen et al., 2009), a prosody assessment method for diagnosis and remediation of speech and language disorders was proposed. A high correlation between automated and judges' individual scores was achieved but the analyses employed by the system require the utterance text-transcription or phonetic segment boundaries.

Phonetic rules can easily be classified as "correct" or "wrong" according to geographic location. In contrast, there is usually more than one intonation pattern that could be considered as "acceptable" given the utterance transcription (Jia et al., 2008). This is due to the fact that intonation provides information about emotions, intentions and attitudes. As a result, instead of classifying an intonation curve as correct or wrong, it is more sensible to motivate the student to follow a given reference intonation pattern on a target form.

In this paper, an automatic intonation assessment system is proposed based on a top-down scheme without any information about the utterance transcription. The proposed method attempts to dissociate the intonation assessment procedure from the resulting phonetic quality of the student's utterance. Given a reference utterance, the student can listen to it and then repeat it by trying to follow the reference intonation curve that must be imitated. Then, the reference and test utterances are aligned frame-by-frame by using Dynamic Time Warping (DTW). Pitch extraction and post-processing are applied to both utterances. The resulting reference and test pitch contours are transformed to a semitone scale and normalized according to the mean. Then, the trend similarity between reference and test intonation templates is evaluated on a frame-by-frame basis by using the DTW alignment mentioned above. Instead of computing the difference between the reference and testing normalized *F*0 contours on a segment-by-segment basis, the current paper proposes to estimate the correlation between both curves. Finally, syllable stress is assessed by using the information provided by the intonation curve combined with frame energy.

The proposed system is not text-dependent (i.e. the text transcription of the reference utterance is not required), minimizes the effect of the resulting phonetic quality in the student's utterance and provides an averaged subjective–objective score correlation (computed as the correlation of human and machine scores) as high as 0.88 when assessing intonation contours. The word stress evaluation system that results from the combination of intonation

and energy contour estimation provides an Equal Error Rate (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 language-independent. The contribution of the paper concerns: (a) a discussion of the role of intonation in second language learning; (b) a text-independent system to evaluate intonation in second language learning; (c) the use of correlation to compare intonation curves as a pattern recognition problem; (d) a text-independent system to assess word stress in second language learning; and, (e) an evaluation of the DTW alignment robustness with respect to the speaker, pronunciation of segments and microphone mismatching conditions.

2. The importance of intonation in second language learning

2.1. Definitions

An adequate phonetic description would be incomplete and unsatisfactory if it does not account for some characteristics accompanying segments that have a relevant meaningful importance. These features are known as suprasegmental elements. The most important ones are pitch, loudness and length (Cruttenden, 2008, pp. 21–23). According to this author: pitch is the perception of fundamental frequency, the acoustic manifestation of intonation; "what is 'loudness' at the receiving end should be related to intensity at the production stage, which in turn is related to the size or amplitude of the vibration"; and, length is related to duration, although "variations of duration in acoustic terms may not correspond to our linguistic judgements of length".

2.2. Intonation

Following (Botinis et al., 2001), "Intonation is defined as the combination of tonal features into larger structural units associated with the acoustic parameter of voice fundamental frequency or *F*0 and its distinctive variations in the speech process. *F*0 is defined by the quasiperiodic number of cycles per second of the speech signal and is measured in Hz". In fact, *F*0 corresponds to the number of times per second that the vocal folds finish a cycle of vibration. Consequently, the production of intonation is regulated by the larynx muscular forces that control the vocal folds tension in addition to aerodynamic forces of the respiratory system. The perceived pitch, which approximately corresponds to *F*0, defines intonation perception.

Intonation has many relevant pragmatic functions that deserve consideration (Chun, 2002; Pierrehumbert and Hirschberg, 1990). At this point it is necessary to state that it is always accompanied by other suprasegmental features, intensity and length, in particular. Among its many functions, it can be said that intonation is particularly relevant to express attitude, prominence, grammatical relationship, discourse structure and naturalness (Roach, 2008; Cruttenden, 2008; Wells, 2006).

Emotions and attitudes are reflected by the intonation that people use when they speak. The same sentence may show different attitudes depending on the intonation with which it is uttered. This is the attitudinal or expressive function of intonation. Additionally, it has a significant role in assigning prominence to syllables that must be recognized as accented. This function is usually called "accentual". Intonation has also a grammatical function as it provides information that makes it easier for the listener to recognize the grammatical and syntactic structure of what is being said, such as determining the placement of phrase, clause or sentence boundary, or the distinction between interrogative and affirmative constructions. This function is commonly referred to as grammatical. Considering the act of speaking from a wider perspective, intonation may suggest to the listener what has to be taken as "new" information and what is considered as "given" information; it may also suggest that the speaker is indicating a kind of contrast or link with some material present in another tone unit and, in conversation, it may provide a hint in relation to the type of answer that is expected. This is the discourse function of intonation. The last function is difficult to describe but is recognizable by every competent native speaker. It has to do with the result of adequate intonation use, which provides naturalness to speech that can be related to the indexical function defined in (Wells, 2006) when he says: "... intonation may act as a marker of personal or social identity. What makes mothers sound like mothers, lovers sound like lovers, lawyers sound like lawyers, ...". Native speaker competence makes it possible to recognize that an utterance has been produced by a native speaker or not. There are many features contributing towards this goal, some of which are more easily distinguishable than others: word choice; syntactic structure; segmental features; and, most certainly, intonation. However competent a foreign speaker may be, if his/her intonation is not the one a native speaker would have used in the same circumstances, his/her speech would sound unnatural and would attract attention to the way he/she said something and not to its contents.

2.3. Stress

Some authors avoid the use of word "stress" because, as mentioned in (Cruttenden, 2008, p. 23), in phonetics and linguistics it is employed in diverse and unclear ways: it is sometimes employed as an equivalent to loudness; sometimes as meaning "made prominent by means other than pitch" (i.e. by intensity or length); and, occasionally, it refers to syllables in lexical items indicating that they have the potential for accent. In this paper the definition presented in (Wells, 2006, p. 3) is followed: "stress is realized by a combination of loudness, pitch and duration".

In a word like "mother", stress falls on the first syllable. In "university", the syllable "ver" receives the *main* or

primary stress, while "u" receives a *secondary* stress. The other syllables, "ni", "si" and "ty" are considered *unstressed*. 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. The importance of secondary stress lies on the fact that in many languages other than English (i.e. Italian and Spanish) it does not affect the pronunciation of segments, as it does in English, where vowel reduction is the result of unstressing some syllables. However, it is common practice to focus the attention on primary stress in second language learning (Jenkins, 2000) as misplacing it affects lexical meaning. Secondary stress misplacing may affect the pronunciation of segments but not necessarily referential meaning. Moreover, due to feasibility issues, the target words in the experiments were chosen in order to avoid secondary stress. Despite the fact that secondary stress is a relevant topic in language acquisition at advanced levels, this research was focused on primary stress. Assessing both types of stress is considered out of the scope of the contribution provided by this paper.

2.4. The importance of Intonation

2.4.1. The importance of intonation in general

As it has been stated in this paper, prosody is significant. Intonation is central in the communication process (Bolinger, 1986, p. 195; Garn-Nunn et al., 1992, p. 107). Speakers of every language recognize this role when they make comments like: "He agreed, but he said it in such a way..." In many occasions the "way" you say something is more important than the literal message, its syntactical organization or the words used to structure it (Fónagy, 2001, p. 583). More frequently than it can be imagined, prosodic features may suggest exactly the opposite meaning than the actual words used by the speaker. Intonation is so significant that it can even be used without a word. A single sound, let us say /m/, can be said with different tones indicating agreement, doubt, disagreement, pleasure, criticism, among other attitudes (Bell, 2009, pp. 1825–1836; Bolinger, 1989, p. 435; Guy and Vonwiller, 1984, pp. 1–17). Not surprisingly it is one the first aspects of speech that children pay attention to, react to, and produce themselves. According to (Peters, 1977), quoted by Cruttenden (2008, p. 291), "Many babies are excellent mimics of intonation and may produce English-sounding intonation patterns on nonsense syllables in the late stages of their pre-linguistic babbling". Besides, there is a close connection between prosody and syntax. As mentioned in (Wells, 2006, p. 11), "Intonation helps identify grammatical structures in speech, rather as punctuation does in writing".

2.4.2. The importance of intonation in foreign language learning

Even though people talk about the intonation of different languages as if they were discrete entities, there are multiple intonation systems within each of these (Grabe and Post, 2002; Fletcher et al., 2005). A native speaker of any language will very easily, and without any previous training, detect that another native speaker of that language is using a dialect different from his/her own, recognizing intonation patterns that are not familiar to him/her. According to (Face, 2006), "With Spanish spoken in different regions of the world, there are considerable differences between the intonation patterns found across the Spanish-speaking world. Even within a relatively small geographic area there can be considerable intonational differences". For instance, to aim at comparing English and Spanish intonation is an impossible task. What might be intended is to compare the intonation of a certain dialect of one of these languages with the intonation of a dialect of the other.

In spite of the fact that there are intonational differences within a language, there are some characteristics that are shared by many languages. As mentioned in (Wells, 2006), "Like other prosodic characteristics, intonation is partly universal, but also partly language-specific". Thus, in many languages a falling tune is associated with a declarative statement or an order, and a rising tune, with an incomplete statement, a question or a polite request. Nevertheless, there are differences that might lead to misunderstanding, particularly of the intentions or attitude of the speaker, who may sound rude or insistent instead of polite, for instance. There is empirical evidence that shows that there are significant differences in the choice of the tone and pitch accent by non-native and native English speakers in similar contexts, which may cause communicative misunderstanding (Ramírez and Romero, 2005). But even though a foreign speaker might use the correct intonation, the problem might lie on the fact that the nucleus is misplaced, where nucleus corresponds to the syllable identified by the final pitch accent (Cruttenden, 2008, p. 271). It is well known that in languages such as French, Italian and Spanish the nucleus is on the last word in the intonational phrase, what is not necessarily the case in English. Consequently, mistakes such as stressing "it" instead of "thought" in "I haven't thought about it", are frequently heard (Cruttenden, 2008, p. 292; Wells, 2006, p. 12). While native English speakers can easily distinguish the grammatical, lexical and pronunciation deviances produced by non-native speakers, and consequently make allowances for their errors, they are incapable to do so for intonation. Following (Wells, 2006, p. 2), "Native speakers of English know that learners have difficulty with vowels and consonants. When interacting with someone who is not a native speaker of English, they make allowances for segmental errors, but they do not make allowances for errors of intonation. This is probably because they do not realize that intonation can be erroneous".

Traditional linguistics has expanded its field from sounds, words, and sentences to larger units, such as full texts, discourses, and interactions, giving rise to disciplines such as discourse analysis, text linguistics, pragmatics, and conversation analysis (Kachru, 1985, p. 2; Celce-Murcia

and Olshtain, 2000, p. 130). It can be said that at present applied linguists stress the crucial importance of intonation, together with stress and rhythm, as their use does not only complement meaning but creates it (Chun, 2002, p. 109; Cruttenden, 2008, p. 328; Morley, 1991, p. 494; Raman, 2004, p. 27). For this reason, the emphasis of present day language teaching is put on communicative effectiveness and, consequently, greater importance in the teaching programme has to be placed on "suprasegmental features rather than on individual sounds" (Morley, 1991, p. 494). In other words there is a tendency to adopt a top-down approach, i.e., to concentrate more on communication and global meaning rather than stick to the traditional bottom-up approach (centred on isolated or contrasted sounds) (Pennington, 1989, pp. 20–38; Dalton and Seidlhofer, 1994, p. 69; Carter and Nunan, 2001, p. 61; Jones, 1997, p. 178). However, it is worth mentioning that the superiority of the top-down over the bottom-up scheme, or vice verse, is still a matter of debate in the field.

3. The proposed system

The system attempts to decide, on a top-down basis, if two utterances (i.e. reference and testing ones), from different speakers, were produced with the same intonation pattern. Fig. 1 shows the block diagram of the proposed scheme to assess the intonation curve generated by a student of a second language. First, $F0$ 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 to allow the comparison of intonation curves from different speakers (e.g. a male and a female). Then, $F0$ contours are smoothed to remove artifacts from the pitch estimation. Then both sequences of MFCC parameters are aligned by using a standard DTW alignment. Finally, the reference and testing $F0$ curves are compared on a frame-by-frame basis by employing the DTW alignment obtained with the MFCC observation sequences. However, rather than estimating the difference between the reference and testing normalized $F0$ patterns on a segment-by-segment basis, the current paper proposes to compute the correlation between both curves. As a result, the reference and testing utterances are compared from the falling-rising trend point of view. In addition, Fig. 2 shows the block diagram of the proposed stress assessment system. In contrast to the intonation assessment method, the stress evaluation system compares the reference and testing templates by employing both $F0$ and energy contours. As it is explained above, stress is the result of the combination of loudness, pitch and duration (Wells, 2006). If pitch is the perception of $F0$, loudness is the perception of signal energy. Consequently, both $F0$ and energy should provide a more accurate assessment of stress than $F0$ or energy individually.

3.1. The intonation assessment system

3.1.1. Pre-processing

First, the speech signals are sampled at 16 kHz and endpoint detected to eliminate silences at the beginning and the ending of each utterance. Then, a high-pass filter at 75 Hz cutoff frequency is applied to reduce the power supply noise. Finally, a pre-emphasis is applied by mean of FIR filter $H(z) = 1 + 0.97z^{-1}$. Observe that the alignment technique between reference and testing utterances uses Mel-frequency cepstral coefficients, and the pre-filtering attempts to equalize the effect of high frequency with respect to low frequency components.

3.1.2. F0 contour extraction and post-processing

After pre-processing, speech signals are low pass filtered at 600 Hz cutoff frequency to eliminate frequencies out of the range of interest and divided into 400-sample frames with a 50% overlap. Then, $F0$ is estimated at each frame and represented in a semitone scale according to:

$$F0_{semitone}(t) = 12\frac{\ln[F0(t)]}{\ln 2}, \quad (1)$$

where $F0(t)$ and $F0_{semitone}(t)$ are, respectively, the fundamental frequency in Hertz and in the semitone scale adopted here at frame t. The logarithm attempts to represent $F0(t)$ according to the human-like perception scale. To reduce doubling or halving errors in $F0$ estimation, curve $F0_{semitone}(t)$ is smoothed according to (Zhao et al., 2007)

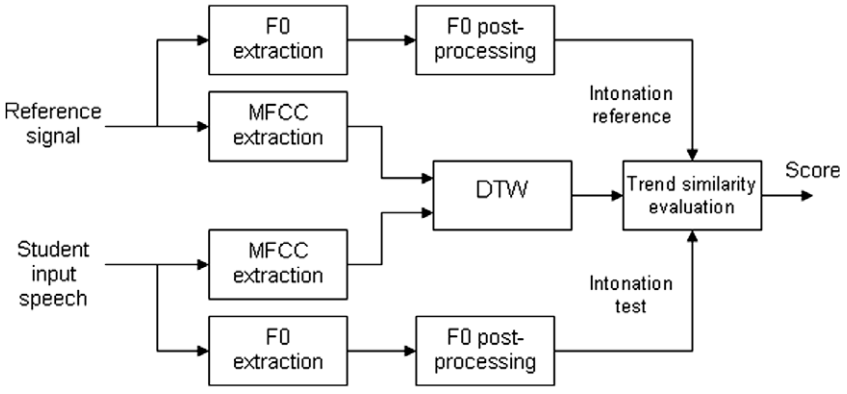

Fig. 1. Block diagram of the proposed intonation assessment system.

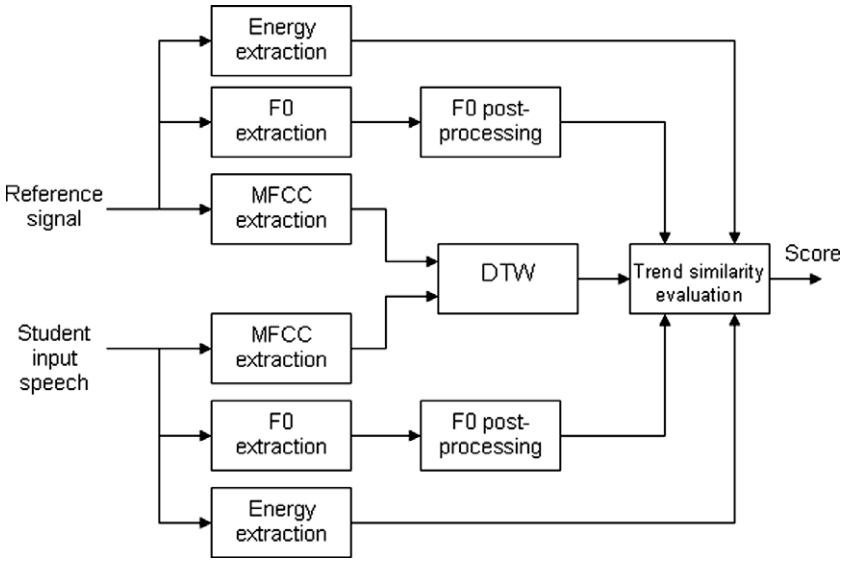

Fig. 2. Block diagram of the proposed stress assessment system.

and with a median filter. Then it is normalized with respect to the mean value. In contrast to (Peabody and Seneff, 2006) where *F*0 contours are normalized with respect to an entire corpus, this paper proposes an utterance based normalization on a top-down scheme. Observe that the intonation patterns in both testing and reference utterances are compared directly without the need of any transcription or predefined correct *F*0 contour shapes. Finally, the discontinuities caused by unvoiced intervals are filled by linear interpolation. The resulting post-processed intonation curve is denoted by $F0_{post\text{-}proc}(t)$.

3.1.3. DTW based alignment

Thirty-three MFCC parameters per frame were 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 Euclidean or Mahalanobis metric. Mahalanobis distance, $d_{mahalanobis}$, is given by:

$$d_{mahalanobis}\left(O^R_{t_1}, O^S_{t_2}\right) = \left[\left(O^R_{t_1} - O^S_{t_2}\right)^T \sum^{-1} \left(O^R_{t_1} - O^S_{t_2}\right)\right]^{1/2}, \quad (2)$$

where O^R_t and O^S_t denote observation vectors in frame t from the reference and testing (student) utterances, respectively; and, Σ is the covariance matrix of the reference and testing utterances. In contrast to the heuristic alignment approach proposed by Delmonte et al. (1997), the dynamic programming method presented here is a structured well-known approach that requires no rules, imposes no bound to the number of features employed in the optimal alignment estimation and requires no text transcription of the reference utterance.

The resulting optimal alignment provided by DTW is indicated by $I(k) = \{i_R(k), i_S(k)\}$, $1 \leqslant k \leqslant K$ where $i_R(k)$ and $i_S(k)$ are the index of frames from the reference and testing utterance, respectively, which are aligned.

Generally, robustness is a key issue in speech processing. Particularly, the massive deployment of speech processing in CALL applications requires robustness to speaker and microphone mismatch. Related to speaker mismatch, different levels of proficiency in the pronunciation of segments may also generate a source of mismatching. Moreover, in this context, the use of different types of low cost microphones is a requirement. As a consequence, several of the experiments presented here attempt to assess the robustness of the proposed approach, besides its accuracy. As it is well known in the literature, the accuracy of DTW-based speech recognition systems is dramatically degraded when the speaker (Rabiner, 1978; Rabiner and Wilpon, 1979; Rabiner and Schmidt, 1980) or channel (Furui, 1981) training–testing 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 or speaker recognition systems. As shown here, speaker and microphone mismatch conditions have a restricted effect in the optimal alignment and in the overall system accuracy.

3.1.4. F0 similarity assessment

In contrast to *F*0 contour classification like the one discussed in (Peabody and Seneff, 2006) to correct tone production in non-native Mandarin, this paper proposes an intonation assessment system that attempts to measure the trend similarity between the intonation curve produced by a student and a reference one. Observe that in Mandarin there are a well-defined number of lexical tones (Tao,

1996). As a consequence, the problem addressed here is not a common topic in pattern classification. According to Fig. 1, the trend similarity between the reference and testing post-processed intonation curves, $F0^R_{post\text{-}proc}(t)$ and $F0^S_{post\text{-}proc}(t)$, respectively, is estimated. As described above, the comparison of both intonation curves is done on a frame-by-frame basis using DTW alignment. However, instead of just estimating the accumulated distance between $F0^R_{post\text{-}proc}(t)$ and $F0^S_{post\text{-}proc}(t)$, this paper proposes that both curves should be compared from the falling-rising trend point of view. In other words, the system should decide if the student was able to produce an intonation curve with the same falling-rising pattern as the reference utterance. Given the DTW alignment between the reference and testing utterances, $I(k)$, mentioned above, the trend similarity measure between both intonation curves, $TS(F0^R_{post\text{-}proc}, F0^S_{post\text{-}proc})$, is defined as the correlation between $F0^R_{post\text{-}proc}$ and $F0^S_{post\text{-}proc}$:

$$TS\left(F0^R_{post\text{-}proc}, F0^S_{post\text{-}proc}\right) = \frac{\sum_{k=1}^{T}\left\{F0^R_{post\text{-}proc}[i_R(k)] - \overline{F0^R_{post\text{-}proc}}\right\}\left\{F0^S_{post\text{-}proc}[i_S(k)] - \overline{F0^S_{post\text{-}proc}}\right\}}{\sigma_{F0^R_{post\text{-}proc}} \cdot \sigma_{F0^S_{post\text{-}proc}}}, \tag{3}$$

where $\sigma_{FO^R_{post\text{-}porc}}$ and $\sigma_{FO^S_{post\text{-}porc}}$ are the standard deviation of $F0^R_{post\text{-}proc}$ and $F0^S_{post\text{-}proc}$, respectively. Alternatively, the trend similarity was also evaluated by using the Euclidean distance between $F0^R_{post\text{-}proc}[i_R(k)]$ and $F0^S_{post\text{-}proc}[i_S(k)]$:

$$TS(F0^R_{post\text{-}proc}, F0^S_{post\text{-}proc}) = \sqrt{\sum_{k=1}^{T}\left\{F0^R_{post\text{-}proc}[i_R(k)] - F0^S_{post\text{-}proc}[i_S(k)]\right\}^2}. \tag{4}$$

Finally, the trend similarity measure between $\frac{dF0^R_{post\text{-}proc}[i_R(k)]}{di_R(k)}$ and $\frac{dF0^R_{post\text{-}proc}[i_S(k)]}{di_S(k)}$ with both correlation and Euclidian distance as trend similarity measures were also considered for comparison purposes:

$$TS\left\{\frac{dF0^R_{post\text{-}proc}[i_R(k)]}{di_R(k)}, \frac{dF0^S_{post\text{-}proc}[i_S(k)]}{di_S(k)}\right\} = \frac{\sum_{k=1}^{K}\left\{\frac{dF0^R_{post\text{-}proc}[i_R(k)]}{di_R(k)} - \overline{\frac{dF0^R_{post\text{-}proc}}{di_R}}\right\}\left\{\frac{dF0^S_{post\text{-}proc}[i_S(k)]}{di_S(k)} - \overline{\frac{dF0^S_{post\text{-}proc}}{di_S}}\right\}}{\sigma_{F0^R_{post\text{-}proc}} \cdot \sigma_{F0^S_{post\text{-}proc}}}, \tag{5}$$

$$TS\left\{\frac{dF0^R_{post\text{-}proc}[i_R(k)]}{di_R(k)}, \frac{dF0^S_{post\text{-}proc}[i_S(k)]}{di_S(k)}\right\} = \sqrt{\sum_{k=1}^{K}\left\{\frac{dF0^R_{post\text{-}proc}[i_R(k)]}{di_R(k)} - \frac{dF0^R_{post\text{-}proc}[i_S(k)]}{di_S(k)}\right\}^2}, \tag{6}$$

where:

$$\frac{dF0^R_{post\text{-}proc}(i_R)}{di_R} = \begin{cases} F0^R_{post\text{-}proc}(i_R) - F0^R_{post\text{-}proc}(i_R - 1) & \text{if } i_R > 0 \\ F0^R_{post\text{-}proc}(1) & \text{if } i_R = 0, \end{cases} \tag{7}$$

$$\frac{dF0^S_{post\text{-}proc}(i_S)}{di_S} = \begin{cases} F0^S_{post\text{-}proc}(i_S) - F0^S_{post\text{-}proc}(i_R - 1) & \text{if } i_S > 0 \\ F0^S_{post\text{-}proc}(1) & \text{if } i_S = 0. \end{cases} \tag{8}$$

The motivation to use the derivative of $F0^R_{post\text{-}proc}$ and $F0^S_{post\text{-}proc}$ instead of the static representation of the curves is due to the fact that the former could represent better the falling-rising trend of the pitch contour that needs to be evaluated.

The proposed intonation assessment system presented here aims at classifying intonation curves according to four patterns that are widely used in the field of linguistics (Wells, 2006, p. 15; Cruttenden, 2008, pp. 271–275; Roach, 2008, p. x): high rise (HR), high fall (HF), low rise (LR) and low fall (LF). The patterns fall-rise and rise-fall are not considered as they are combinations of the four basic ones mentioned above. As described in Section 2.2, intonation has many functions that are not univocally related to any of the patterns addressed here. Consequently, a detailed discussion on the functionality of the intonation reference models is out of the scope of the current paper by definition.

3.2. The stress assessment system

The stress evaluation system, which is represented in Fig. 2, is generated from the scheme in Fig. 1. The energy (intensity) extraction contour is included and combined with the post-processed intonation 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:

$$E(t) = 10 \cdot \log\left[\sum_{n=1}^{N} x^2(t+n)\right], \tag{9}$$

where $x(\cdot)$ denotes the signal samples and N is the frame width. If $E^R(t)$ and $E^S(t)$ denote the energy contour of the reference and testing utterances, respectively, the trend similarity that includes the intonation and energy contour, $TS\left(F0^R_{post\text{-}proc}, E^R, F0^S_{post\text{-}proc}, E^S\right)$, is computed as:

$$TS\left(F0^R_{post\text{-}proc}, E^R, F0^S_{post\text{-}proc}, E^S\right) = \alpha \cdot TS(E^R, E^S) + (1-\alpha) \cdot TS\left(F0^R_{post\text{-}proc}, F0^S_{post\text{-}proc}\right), \tag{10}$$

where: $TS(E^R, E^S)$ and $TS(F0^R_{post\text{-}proc}, F0^S_{post\text{-}proc})$ are estimated according to (3) by making use of the correlation between E^R and E^S, and between $F0^R_{post\text{-}proc}$ and $F0^S_{post\text{-}proc}$, respectively; and, α is a weighting factor. Finally, the system takes the decision about the stress pattern resulted from the student's utterance, SD, according to:

$$SD\left[TS\left(F0^{R}_{post\text{-}proc}, E^{R}, F0^{S}_{post\text{-}proc}, E^{S}\right)\right] = \begin{cases} \text{the same as the reference} & \text{if} \quad TS\left(F0^{R}_{post\text{-}proc}, E^{R}, F0^{S}_{post\text{-}proc}, E^{S}\right) \geqslant \theta_{SD} \\ \text{different from the reference} & \text{elsewhere,} \end{cases} \tag{11}$$

where θ_{SD} correspond to a decision threshold, which in turn depends on the target false positive and false negative rates.

4. Experiments

4.1. Databases

Two databases were recorded at the Speech Processing and Transmission Laboratory (LPTV), Universidad de Chile, to evaluate the performance of the proposed schemes to address the problems of intonation and stress assessment. All the speech material was recorded in an office environment with a sampling frequency equal to 16 kHz. There were 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). The databases are described as follows.

4.1.1. Intonation assessment data set

In order to avoid additional difficulties from the user point of view, short sentences that do not include uncommon words or complicated syntactic structures were chosen. They use the most usual intonation patterns: HR, HF, LR, and LF. Observe that in the testing procedure, the students are expected to reproduce the intonation patterns following the model sentences heard, contrasting their realizations with the reference utterance. This data set is composed of six sentences: "What's your name"; "My name is Peter"; "It's made of wood"; "It's terrible"; "It was too expensive"; and, "I tried both methods". The sentences were uttered with the intonation patterns mentioned above: HR, HF, LR and LF. Altogether there are 6 sentences $\times$ 4 intonation patterns = 24 types of utterances that were recorded by 16 speakers (eight experts and eight non-experts in English language and phonetics) by making use of three microphones simultaneously. Then, the total number of recorded sentences is equal to 24 types of utterances $\times$ 16 speakers $\times$ 3 microphones = 1552 utterances. In the experiment of intonation assessment, the reference utterances correspond to the sentences recorded by one of the experts in English language and phonetics (the most senior one). The number of possible experiments per target sentence per speaker per microphone is equal to 4 reference intonation pattern labels $\times$ 4 testing intonation pattern labels = 16 experiments. Finally, the total number of intonation assessment experiments is equal to 16 experiments per speaker per sentence per microphone $\times$ 15 testing speakers $\times$ 6 types of sentences $\times$ 3 microphones = 4320 experiments.

4.1.2. Stress assessment data set

Firstly, due to feasibility issues, the target words were chosen in order to avoid secondary stress. Despite the fact that secondary stress is a relevant topic in language acquisition as it may affect the pronunciation of segments, this research focused on primary stress, the misplacing of which may affect referential meaning. Assessing both types of stress was considered out of the scope of the contribution provided by the current paper. In this context, the selected words are composed of two, three and four syllables. For each case, four examples were generated. 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 number of stress variants is equal to the number of syllables in the target word. Consequently, altogether there are 4 words $\times$ (2 syllables + 3 syllables + 4 syllables) = 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.

4.2. Experimental set-up

The DTW algorithm mentioned in Figs. 1 and 2 was implemented according to (Sakoe and Chiba, 1978). The covariance matrix employed by Mahalanobis distance in (2) was estimated with a subset of the intonation assessment database explained in Section 4.1.1. The fundamental frequency $F0$ is estimated by using the autocorrelation based Praat pitch detector system (Boersma and Weenink, 2008). As mentioned above, the utterances are divided into 400-sample frames with a 50% overlap. Thirty-three MFCC parameters per frame were computed: the frame energy plus ten static coefficients and their first and second time derivatives.

4.3. Subjective–objective score correlation

The subjective–objective score correlation is estimated as the correlation between the subjective scores and the objective scores delivered by the automatic intonation assessment system proposed. The subjective scores are generated according to the procedure described as follows. First, an expert in phonetics and English language (the most senior one) recorded all the sentences with all the intonation patterns described in Section 4.1.1. These utterances were selected as reference and each one was labelled with HR, HF, LR, or LF (see Section 3.1.4). Then, the remaining seven expert speakers listened to and repeated each reference utterance by following the corresponding intonation pattern. In the same way the eight non-expert speakers recorded the reference utterances, but they were supervised by the seven experts to make sure that the intonation pattern was reproduced correctly. Then, the utterances recorded by the seven expert and the eight non-expert speakers were also labelled with HR, HF, LR, or LF. Finally, an engineer checked the concordance between the utterances and the assigned intonation pattern label. Most of the papers in the field of CAPT (Computer Aided Pronunciation Training) employ the subjective–objective score correlation to evaluate the accuracy of a given system. In this context, Tables 1 and 2 define the subjective scores when a student testing utterance is compared with a reference one that contains the intonation pattern to be followed. Accordingly, the subjective scores, that result from the direct comparison between reference and testing intonation pattern labels, are defined in Tables 1 and 2. Consider that $SubjEvaluation_{Testing}$ and $SubjEvaluation_{Reference}$ denote the subjective evaluation in the testing and reference utterances, respectively, where $SubjEvaluation_{Testing}$ and $SubjEvaluation_{Reference}$ are one of the following categories regarding the intonation pattern: HF; LF; HR; and, LR. Therefore, the strict subjective score (Table 1) that results from the comparison of the testing and reference intonation patterns are defined as follows:

$$\text{Strict subjective score} = \begin{cases} 5 & \text{if} \quad SubjEvaluation_{Testing} = SubjEvaluation_{\text{Reference}} \\ 1 & \text{elsewhere.} \end{cases} \tag{12}$$

Accordingly, Table 2 defines the non-strict subjective scores as follows:

$$\text{Non-strict subjective score} = \begin{cases} 5 & \text{if} \quad SubjEvaluation_{Testing} = SubjEvaluation_{\text{Reference}} \\ 4 & \text{if} \quad (SubjEvaluation_{Testing}, SubjEvaluation_{Testing}) \in \{(HF, LF), (LF, HF), (HR, LR), (LR, HR)\} \\ 1 & \text{elsewhere.} \end{cases} \tag{13}$$

As shown in (13), HF/LF and HR/LR substitutions were labelled with score 4 because score 3 is neutral and score 2 is negative. It sounds sensible to provide a positive score if the student reproduced an intonation pattern similar to the reference one, although not exactly the same.

4.4. DTW alignment accuracy experiments

As mentioned above, the speaker, pronunciation of segments and microphone mismatch effect on DTW accuracy alignment is evaluated in this paper. A subset of three expert speakers and two non-expert speakers from the intonation data set (Section 4.1.1) were selected to assess the robustness 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 here as:

$$E_{align}(b) = 100 \cdot \frac{d(b)}{D}, \tag{14}$$

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

$$d(b) = \frac{1}{2}\sqrt{d_R(b)^2 + d_S(b)^2}, \tag{15}$$

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. 3). Given two utterances with the same text transcription, the total alignment error, E_{align}, is equal to:

$$E_{align} = \frac{1}{B}\sum_{b=1}^{B} E_{align}(b), \tag{16}$$

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

5. Results and discussion

5.1. Alignment experiments

Table 3 shows the DTW alignment error using several features in combination with Euclidian distance. The data set explained in Section 4.1.1 was employed. All the utterances with the same transcription were compared two-by-two independently of the speaker and microphone matching condition. As can be seen in Table 3, the lowest alignment

Table 1
Strict subjective score scale criterion for intonation contour comparison defined as in Section 4.3. HF, LF, HR and LR denote, respectively, high fall, low fall, high rise and low rise as defined in Section 4.1.

Subjective intonation pattern label in the testing utterance ($SubjEvaluation_{Testing}$)	Subjective intonation pattern label in the reference utterance ($SubjEvaluation_{Reference}$)			
	HF	LF	HR	LR
HF	5	1	1	1
LF	1	5	1	1
HR	1	1	5	1
LR	1	1	1	5

Table 2
Non-strict subjective score scale criterion for intonation contour comparison defined as in Section 4.3. HF, LF, HR and LR denote, respectively, high fall, low fall, high rise and low rise as defined in Section 4.1.

Subjective intonation pattern label in the testing utterance ($SubjEvaluation_{Testing}$)	Subjective intonation pattern label in the reference utterance ($SubjEvaluation_{Reference}$)			
	HF	LF	HR	LR
HF	5	4	1	1
LF	4	5	1	1
HR	1	1	5	4
LR	1	1	4	5

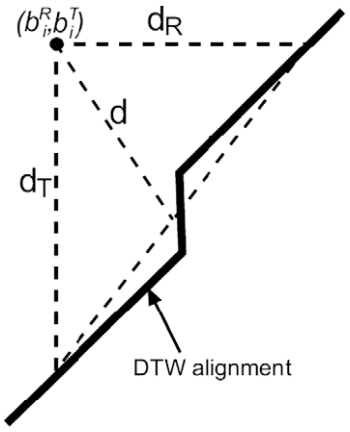

Fig. 3. Representation of DTW alignment error measure, d. Point (b_i^R, b_i^S) indicates the intersection of boundary i within the reference and testing utterances. The distances d_R and d_T are the horizontal and vertical distances, respectively, between the phonetic boundaries and the DTW alignment.

error takes places with MFCC features in combination with frame energy (statistically significant with $p < 0.0001$ when compared with the other features combinations). Table 4 compares the DTW alignment error between speaker matched and unmatched condition, where both Euclidean and Mahalanobis distance were employed in combination with MFCC plus energy. When the Euclidean metric is replaced with the Mahalanobis distance, the error is reduced by 10% (this difference is statistically significant with $p < 0.0001$). Also in Table 4, when compared with

Table 3
Alignment error by using different features in DTW. Local distance corresponds to the Euclidian metric. The sample size is equal to 5 speakers × 4 intonation patterns × 2 microphones = 40 utterances per target sentence, which in turn generates 780 pair combinations per target sentence. Considering 6 target sentences as explained in Section 4.1.1, there are altogether 780 pair combinations per target sentence × 6 target sentences = 4680 experiments.

Feature	Alignment error (%)
Frame energy	13.27
$F0$	11.49
$F0$ + frame energy	11.06
MFCC	5.31
MFCC + frame energy	4.90

Table 4
Alignment error with speaker matched and unmatched condition. The sample size is equal to 5 speakers × 4 intonation patterns × 2 microphones = 40 utterances per target sentence, which in turn generates 780 pair combinations per target sentence. Considering 6 target sentences as explained in Section 4.1.1, there are altogether 780 pair combinations per target sentence × 6 target sentences = 4680 experiments.

Speaker matching condition	Euclidean distance (%)	Mahalanobis distance (%)
Matched	3.10	2.86
Unmatched	4.78	4.22

Table 5
Alignment error with microphone matched and unmatched condition. The sample size is equal to 5 speakers × 4 intonation patterns × 2 microphones = 40 utterances per target sentence, which in turn generates 780 pair combinations per target sentence. Considering 6 target sentences as explained in Section 4.1.1, there are altogether 780 pair combinations per target sentence × 6 target sentences = 4680 experiments.

Microphone matching condition	Euclidean distance (%)	Mahalanobis distance (%)
Matched	3.10	2.86
Unmatched	3.22	2.89

speaker matching condition, the alignment error shows an increase of just 1.68% and 1.36% points when utterances are from different speakers with Euclidean and Mahalanobis distances, respectively. Consequently, this result suggests that the DTW alignment is robust to speaker mismatch.

Table 5 shows alignment error between different matched and mismatched microphone conditions between the reference and testing utterances. As can be seen, when compared with microphone matching condition, the alignment error shows an increase of just 0.12% and 0.03% points, when testing and reference utterances are recorded with different microphones, with Euclidean and Mahalanobis distances, respectively. 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 4 and 5 strongly suggest that the DTW alignment is robust to speaker and microphone mismatch.

Table 6
Averaged subjective–objective score correlation in intonation assessment with different trend similarity measures. Strict and non-strict subjective scores are defined in Tables 1 and 2. The number of possible experiments per target sentence per speaker per microphone is equal to 4 reference intonation pattern labels × 4 testing intonation pattern labels = 16 experiments. Finally, the total number of intonation assessment experiments is equal to 16 experiments per speaker per sentence per microphone × 15 testing speakers × 6 types of sentences × 3 microphones = 4320 experiments, as is explained in Section 4.1.1.

Trend similarity measure	Strict subjective score	Non-strict subjective score
Correlation	0.54	0.88
Euclidian distance	0.40	0.62
Correlation (D)	0.48	0.79
Euclidian distance (D)	0.31	0.46

Table 7
Averaged objective–subjective score correlation in intonation contour assessment. Speaker matched and unmatched conditions are compared, by using the non-strict subjective score scale explained in Table 1. The number of possible experiments per target sentence per speaker per microphone is equal to 4 reference intonation pattern labels × 4 testing intonation pattern labels = 16 experiments. Finally, the total number of intonation assessment experiments is equal to 16 experiments per speaker per sentence per microphone × 15 testing speakers × 6 types of sentences × 3 microphones = 4320 experiments, as is explained in Section 4.1.1.

Trend similarity measure	Speaker matched condition	Speaker unmatched condition
Correlation	0.88	0.88
Euclidian distance	0.71	0.62
Correlation (D)	0.79	0.79
Euclidian distance (D)	0.57	0.46

5.2. Intonation experiments

Table 6 shows the averaged subjective–objective score correlation between the trend similarity provided by the system in Fig. 1 and the subjective score within the intonation assessment database mentioned in Section 4.1.1. Strict and non-strict subjective scores between reference and testing utterances are defined in Tables 1 and 2, respectively. According to Table 6, the highest average subjective–objective score correlation is given by the use of correlation as a trend similarity measure (statistically significant with $p < 0.0001$ when compared with the other trend similarity measures). With the non-strict subjective score scale, the averaged subjective–objective score correlation is as high as 0.88. However, with the strict subjective score scale the averaged subjective–objective score correlation is substantially decreased (statistically significant with $p < 0.0001$). This result suggests that the proposed system can accurately discriminate between rising and falling pitch contours. In contrast, the accuracy to distinguish between HF and LF or between HR and LR is reduced.

The DTW alignment robustness to speaker mismatching suggested by Table 4 is corroborated in Table 7, which shows the averaged subjective–objective score correlation in intonation assessment with and without speaker matching condition. As can be seen, speaker unmatched condition leads to a mean reduction in the averaged subjective–objective score correlation as low as 8.2%. Moreover, in the context of second language learning, pronunciation of segments is also a source of mismatch between reference and testing utterances. Table 8 presents the averaged objective–subjective score correlation in intonation assessment with and without pronunciation of segments matching condition. In the former case, the reference and testing utterances come from the experts in English language and phonetics. In the latter case, the testing utterances were pronounced by non-expert in English language speakers. According to Table 8, pronunciation of segments mismatch leads to a reduction in the averaged subjective–objective score correlation in intonation assessment as low as 2.5%; 7.6%; 0.5%; and, 0% with trend similarity estimated with (3)–(6), respectively. This result also suggests the validity of the hypothesis concerning the DTW alignment robustness to speaker and pronunciation quality mismatch.

Table 8
Averaged objective–subjective score correlation in intonation contour assessment. Expert speaker matched and unmatched conditions are compared, by using the non-strict subjective score scale explained in Table 2. The number of possible experiments per target sentence per speaker per microphone is equal to 4 reference intonation pattern labels × 4 testing intonation pattern labels = 16 experiments. Finally, the total number of intonation assessment experiments is equal to 16 experiments per speaker per sentence per microphone × 15 testing speakers × 6 types of sentences × 3 microphones = 4320 experiments, as is explained in Section 4.1.1.

Trend similarity measure	Expert speakers (pronunciation of segments matching condition)	No-experts speakers (pronunciation of segments mismatching condition)
Correlation	0.89	0.87
Euclidian distance	0.65	0.60
Correlation (D)	0.79	0.79
Euclidian distance (D)	0.44	0.53

Fig. 4 shows the averaged subjective–objective score correlation in intonation assessment with microphone matched and unmatched condition. The reference utterances were recorded with Mic1. The testing utterances were captured with Mic1, Mic2 and Mic3. As can be seen in Fig. 4, the mean difference in averaged subjective–objective score correlation in intonation assessment between matched and unmatched condition is just equal to 2.5%. This result strongly corroborates the result discussed in Table 5.

5.3. Stress experiments

Fig. 5 presents the receiver operating characteristic (ROC) curves (false negative rate, FNR, and false positive rate, FPR) with the stress assessment system shown in Fig. 2. The trend similarity is estimated with (10) and the final decision about stress assessment is taken according to (11). The variable α was tuned in order to minimize

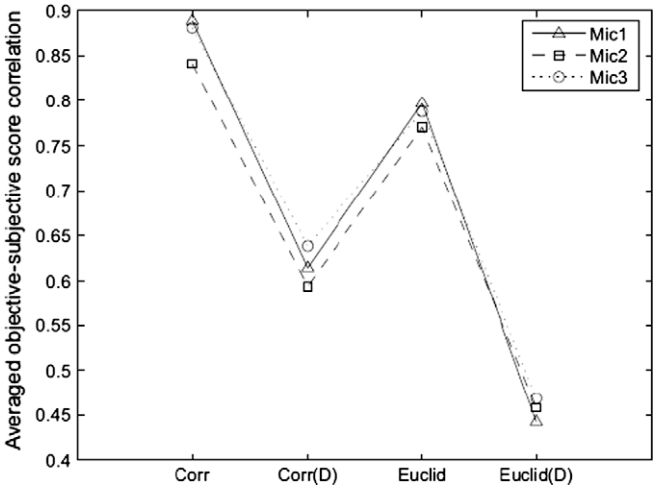

Fig. 4. Averaged objective–subjective score correlation in intonation assessment with different microphones. Mic1 represents the high quality microphone employed. Mic2 and Mic3 represent low-cost desktop PC microphones.

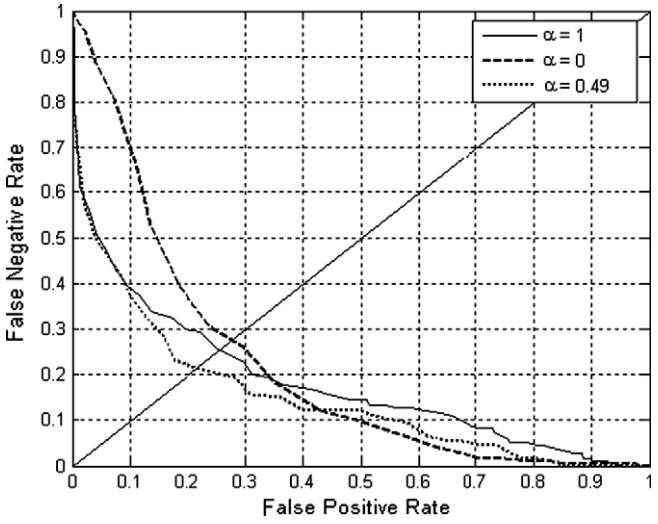

Fig. 5. False negative and false positive ROC curves in stress evaluation. The trend similarity measure is estimated according to (10) and the decision is taken by using (11). $\alpha = 1$ indicates that only pitch contour is employed and $\alpha = 0$ indicates that only frame energy contour is employed.

the area below the ROC curve and the optimal value is equal to 0.49. Fig. 5 also shows the FPR/FNR curves with $\alpha = 0$, $\alpha = 1$ and $\alpha = 0.49$. Table 9 presents the area below the ROC curve and EER with α equal to 0, 1 and 0.49. According to Fig. 5 and Table 9, the optimal α 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 $\alpha = 1$. Significance analysis using McNemar's test (Gillick and Cox, 1989) shows that the differences in EER between $\alpha = 0.49$ and $\alpha = 1$, and between $\alpha = 0.49$ and $\alpha = 0$ are significant with $p < 0.00048$ and $p < 0.077$, respectively. This result suggests that both pitch and energy contours provide relevant information to assess word stress. The stress assessment system accuracy should

Table 9
ROC area and equal error rate (EER) for stress assessment system for different α, using correlation as trend similarity measure. The optimal α that minimizes EER is equal to 0.49. The sample size is equal to 36 experiments per speaker per microphone × 7 testing speakers × 3 microphones = 756 experiments, as is explained in Section 4.1.2.

Feature	ROC area	EER (%)
$\alpha = 1$	0.181	25.41
$\alpha = 0$	0.212	27.64
$\alpha = 0.49$	0.147	21.48

be improved by including duration information, which in turn is not straightforward in the frame of the DTW alignment. However, it is worth mentioning that word based state-of-the-art automatic pronunciation assessment systems provide subjective–objective score correlation between 0.6 and 0.8 depending, among other factors, on the number of levels in the evaluation scale (Dong and Yan, 2008; Tepperman and Narayanan, 2007; Stouten and Martens, 2006; Su et al., 2006; Oppelstrup et al., 2005; Bernstein et al., 1990; Eskenazi, 1996; Hiller et al., 1994). In (Molina et al., 2009), the classification error, defined as the difference between the subjective and objective score levels, was estimated in a word based CAPT system with two and five level scales. With a two level scale, the subjective–objective score correlation is 0.8 in average and the classification error is around 10%. Also, with a five level scale, the subjective–objective score correlation is 0.67 in average and the classification error is around 55%. As a result, the optimal EER provided by the proposed stress assessment system (21.5%) is similar to phonetic pronunciation assessment systems. This suggests that the proposed scheme should be accurate enough for practical applications.

The proposed method requires a reference intonation pattern that the student should try to follow. However, the text transcription of both reference and testing utterances are not required. The motivation behind the proposed strategy, as explained here, is the fact that there is not a clear definition of "correct" or "wrong" intonation (Jia et al., 2008). The same sentence may be pronounced with several intonation patterns according to the context and in most cases there should not be only one correct intonation. The problem addressed by the current paper is how to teach a student to follow a given intonation pattern as a reference provided that there is not only one correct intonation production. In contrast, scoring intonation of spontaneous speech without a reference is out of the scope of the hypothesis considered in this paper.

6. Conclusions

A discussion of the nature and importance of intonation in second language learning is presented in this paper. As a result, a text-independent and language-independent automatic intonation assessment system for second language learning is proposed based on a top-down scheme. A stress

assessment system is also presented by combining intonation and energy contour estimation. The system directly compares the utterance pronounced by the student with a reference one. The trend similarity of intonation 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 microphone, speaker and quality pronunciation mismatch is addressed. The intonation assessment system achieves an averaged subjective–objective score correlation as high as 0.88 when correlation as trend similarity measure is employed. 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 could be employed in real applications. Despite of 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.

Acknowledgements

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

References

Baetens, H., 1982. Bilingualism: basic principles, on-line version.

Bell, N., 2009. Responses to failed humor. J. Pragmatics 41 (9), 1825–1836 (2009).

Bernat, E., 2006. Assessing EAP learners' beliefs about language learning in the Australian context. Asian EFL J. 8 (2) (Article 9).

Bernstein J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., 1990. Automatic evaluation and training in English pronunciation, In: Proc. Internat. Congress on Spoken Language Processing (ICSLP) '90, pp. 1185–1188.

Boersma, P., Weenink, D., 2008. Praat: doing phonetics by computer (Version 5.0.29) [Computer program]. http://www.praat.org/ (Retrieved 14.07.08).

Bolinger, D., 1986. Intonation and its Parts: Melody in Spoken English. Stanford University Press, Stanford.

Bolinger, D., 1989. Intonation and its Uses: Melody in Grammar and Discourse. Stanford University Press, Stanford.

Botinis, A., Granström, B., Möbius, B., 2001. Developments and paradigms in intonation research. Speech Comm. 33 (4), 263–296.

Carter, R., Nunan, D., 2001. The Cambridge Guide to Teaching English to Speakers of Other Languages. Cambridge University Press, Cambridge.

Celce-Murcia, M., Olshtain, E., 2000. Discourse and Context in Language Teaching: A Guide for Language Teachers. Cambridge University Press, Cambridge.

Chun, D., 2002. Discourse Intonation in L2. John Benjamins.

Cruttenden, A., 2008. Gimson's Pronunciation of English, seventh ed. Hodder Education, London.

Dalton, C., Seidlhofer, B., 1994. Pronunciation. Oxford University Press, Oxford.

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.

Dong, B., Yan, Y., 2008. A synchronous method for automatic scoring of language learning. In: Sixth Internat. Symposium on Chinese Spoken Language Processing, ISCSLP '08.

Dong, B., Zhao, Q., Zhang, J., Yan, Y., 2004. Automatic assessment of pronunciation quality. In: Internat. Symposium on Chinese Spoken Language Processing, pp. 137–140.

El-Imam, Y.A., Don, Z.M., 2005. Rules and algorithms for phonetic transcription of standard Malay. IEICE Trans. Inform. Systems.

Eskenazi, M., 1996. Detection of foreign speakers' pronunciation errors for second language training – preliminary results. In: Proc. Internat. Congress on Spoken Language Processing (ICSLP) '96, pp. 1465–1468.

Eskenazi, M., Hansma, S., 1998. The fluency pronunciation trainer. In: Proc. STiLL Workshop on Speech Technology in Language Learning, Marhollmen.

Face, T., 2006. Narrow focus intonation in Castillian Spanish absolute negatives. J. Lang. Linguist. 15 (2), 295–311.

Fletcher, J., Grabe E., Warren, P., 2005. Intonational Variation in Four Dialects of English: The High Rising Tune.

Fónagy, I., 2001. Languages Within Language: An Evolutive Approach. John Benjamins, Amsterdam/Philadelphia.

Franco, H., Neumeyer, L., Kim, Y., Ronen, O., 1997. Automatic pronunciation scoring for language instruction. In: ICASSP'97, Vol. 2. pp. 1471–1474.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29 (2), 254–272.

Garn-Nunn, P., Lynn, J., Calvert, D., 1992. Calvert's Descriptive Phonetics. Thieme Medical Publishers, Inc., New York.

Gillick, L., Cox, S.J., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: Proc. ICASSP'89, Glasgow, Scotland, pp. 532–535.

Grabe, E., Post, B., 2002. Intonational Variation in the British Isles.

Gu, L., Harris, J., 2003. SLAP: a system for the detection and correction of pronunciation for second language acquisition. In: Internat. Symposium on Circuits and Systems, ISCAS '03, Vol. 2. pp. 580–583.

Guy, G., Vonwiller, J., 1984. The meaning of an intonation in Australian English. Australian J. Linguist. 4 (1), 1–17, 1469–2996.

Hiller, S., Rooney, E., Vaughan, R., Eckert, M., Laver, J., Jack, M., 1994. An automated system for computer-aided pronunciation learning. Comput. Assist. Lang. Learn. 7 (1994), 51–63.

Holmes, J., Holmes, W., 2001. Speech Synthesis and Recognition, second ed. CRC Press.

Jenkins, J., 2000. The Phonology of English as an International Language. Oxford University Press.

Jia, H., Tao, J., Wang X., 2008. Prosody variation: application to automatic prosody evaluation of Mandarin speech. In: Proc. Speech Prosody, pp. 547–550.

Jones, R., 1997. Beyond 'listen and repeat': pronunciation teaching materials and theories of second language acquisition. System 25 (1), 103–112.

Jurafsky, D., Martin, J., 2009. Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics, and Speech Recognition, second ed.

Kachru, Y., 1985. Discourse strategies, pragmatics and ESL. Where are we going?. RELC J. 16 (2) 1–17.

Kim, H., Sung, W., 2002. Implementation of an intonational quality assessment system. In: ICSLP-2002, pp. 1225–1228.

Liang, W., Liu, J., Liu, R., 2005. Automatic spoken English test for Chinese learners. In: Proc. International Conference on Communications, Circuits and Systems, Vol. 2, pp. 857-860.

Molina, C., Yoma, N.B., Wuth, J., Vivanco, H., 2009. ASR based pronunciation evaluation with automatically generated competing vocabulary. Speech Comm. 51 (6), 485–498.

Morley, J., 1991. The pronunciation component in teaching English to speakers of other languages. TESOL Quart. 25 (3), 481–520.

Moyer, A., 2004. Age, Accent and Experience in Second Language Acquisition: An Integrated Approach to Critical Period Inquiry. Multilingual Matters, Clevedon.

Neumeyer, L., Franco, H., Weintraub, M., Price, P., 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proc. ICSLP '96.

Oppelstrup, L., Blomberg, M., Elenius, D., 2005. Scoring children's foreign language pronunciation. In: Proc. FONETIK, Goteborg.

Peabody, M., Seneff, S., 2006. Towards automatic tone correction in non-native Mandarin. In: Proc. Fifth Internat. Symposium on Chinese Spoken Language Processing (ISCSLP), Kent Ridge, Singapore.

Pennington, M., 1989. Teaching pronunciation from the top down. RELC J. 20 (1), 20–38.

Peters, A.M., 1977. Language learning strategies: does the whole equal the sum of the parts? Language 53, 560–573.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning intonational contours in English. In: Cohen, P., Morgan, J., Pollack, M. (Eds.), Intentions in Communication. MIT Press.

Rabiner, L., 1978. On creating reference templates for speaker independent recognition of isolated words. IEEE Trans. Acoust. Speech Signal Process. 26 (1), 34–42.

Rabiner, L., Schmidt, C., 1980. Application of dynamic time warping to connected digit recognition. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 377–388.

Rabiner, L., Wilpon, J., 1979. Speaker-independent isolated word recognition for a moderate size (54 word) vocabulary. IEEE Trans. Acoust. Speech Signal Process. 27 (6), 583–587.

Raman, M., 2004. English Language Teaching. Atlantic Publishers and Distributors, New Delhi.

Ramírez, D., Romero, J., 2005. The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers. Intercult. Pragmatics 2, 151–168.

Roach, P., 2008. English Phonetics and Phonology, third ed. Cambridge University Press, Cambridge.

Rypa, M., Price, P., 1999. VILTS: a tale of two technologies. CALICO J. 16 (3), 385–404.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Proc. ASSP-26.

Saussure, F., Bouquet, S., Engler, R., 2006. Writings in General Linguistics. Oxford University Press, Oxford.

Shimizu, M., Taniguchi, M., 2005. Reaffirming the Effect of Interactive Visual Feedback on Teaching English Intonation to Japanese Learners. In: Phonetics Teaching and Learning Conference 2005. University College, London.

Stouten, F., Martens, J.P., 2006. On the use of phonological features for pronunciation scoring. In: Proc. ICASSP.

Su, P., Chen, Q., Wang, X., 2006. A fuzzy pronunciation evaluation model for English learning. In: Proc. Fifth Internat. Conf. on Machine Learning and Cybernetics, Dalian, 13–16 August.

Tao, H., 1996. Units in Mandarin Conversion: Prosody, Discourse and Grammar. John Benjamins.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K., Sonmez, K., 2000. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In: Proc. ICSLP.

Tepperman, J., Narayanan, S., 2007. Using articulatory representations to detect segmental errors in nonnative pronunciation. IEEE Trans. Audio Speech Lang. Process. 16 (1), 8–22.

Tepperman, J., Narayanan, S., 2008. Better nonnative intonation scores through prosodic theory. In: Proc. InterSpeech ICSLP, Brisbane, Australia.

Traynor, P.L., 2003. Effects of computer-assisted-instruction on different learners. Instruct. Psychol. J., 137–143.

van Santen, J.P.H., Prud'hommeaux, E., Black, L.M., 2009. Automated measures for assessment of prosody. Speech Comm. 51 (11), 1082–1097.

Wells, J.C., 2006. English Intonation. Cambridge University Press, Cambridge.

You, K., Kim, H., Sung, W., 2004. Implementation of an intonational quality assessment system for a handheld device. In: INTERSPEECH-2004, pp. 1857–1860.

Zhao, X., O'Shaughnessy, D., Minh-Quang, N., 2007. A processing method for pitch smoothing based on autocorrelation and cepstral F0 detection approaches. In: ISSSE '07, Internat. Symposium on Signals, Systems and Electronics, pp. 59–62.