Multicriteria-Based Computer-Aided Pronunciation Quality Evaluation of Sentences

Néstor Becerra Yoma, Leopoldo Benavides Berrios, Jorge Wuth Sepúlveda, and Hiram Vivanco Torres

The problem of the sentence-based pronunciation evaluation task is defined in the context of subjective criteria. Three subjective criteria (that is, the minimum subjective word score, the mean subjective word score, and first impression) are proposed and modeled with the combination of word-based assessment. Then, the subjective criteria are approximated with objective sentence pronunciation scores obtained with the combination of word-based metrics. No *a priori* studies of common mistakes are required, and class-based language models are used to incorporate incorrect and correct pronunciations. Incorrect pronunciations are automatically incorporated by making use of a competitive lexicon and the phonetic rules of students' mother and target languages. This procedure is applicable to any second language learning context, and subjective-objective sentence score correlations greater than or equal to 0.5 can be achieved when the proposed sentence-based pronunciation criteria are approximated with combinations of word-based scores. Finally, the subjective-objective sentence score correlations reported here are very comparable with those published elsewhere resulting from methods that require *a priori* studies of pronunciation errors.

Keywords: Computer-aided pronunciation training, subjective criteria, second language learning, ASR.

Manuscript received Jan. 7, 2012; revised May 5, 2012; accepted July 16, 2012.

This work was funded by Conicyt-Chile under grant Fondecyt No. 1100195.

Néstor Becerra Yoma (phone: +56 2 978 4205, nbecerra@ing.uchile.cl), Leopoldo Benavides Berrios (lbenavid@gmail.com), and Jorge Wuth Sepúlveda (jwuth@ing.uchile.cl) are with the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile.

Hiram Vivanco Torres (hvivanco@uchile.cl) is with the Department of Linguistics, Universidad de Chile, Santiago, Chile.

http://dx.doi.org/10.4218/etrij.13.0112.0016

I. Introduction

The spread and popularity of computer-aided pronunciation training in second language learning (2LL) certainly depends on the flexibility to apply several evaluation criteria and the efficient generation of didactical material with minimal human supervision. When learning a foreign language, several criteria must be used to evaluate a speaker's proficiency, including the mastery of morphosyntactical, lexical, discoursal, and pragmatic features, among others. In the case of oral production, phonetic considerations must be added, including segmental and suprasegmental characteristics. For example, students of the basic Spanish language courses can be evaluated on oral expression according to the following criteria: communicative success, grammar, vocabulary, and fluency.[1)]

Considering oral expression, it is well known that subjective evaluation of pronunciation is extremely complex, as there are several aspects that require observation, such as intelligibility and allophonic accuracy.

Regarding allophonic evaluation, state-of-the-art automatic speech recognition (ASR)-based computer-aided pronunciation training (CAPT) methods still show limitations and thus lack accuracy. On the other hand, accurate identification of phonetic substance does not ensure the understanding of the message by the listener, who in many cases must rely on extra-phonetic information. Given this context, the automatic evaluation of oral production should rely on units larger than the phoneme and answer such questions about the expression as the following: Is it intelligible? Is it grammatical? Does it make sense? Is it coherent? It might be convenient to holistically

1) http://sip.la.psu.edu/blp/files/criteria_oral.pdf

evaluate the expression, based on first impression, or to evaluate each of the aforementioned aspects separately to obtain an average. It might also be convenient to consider the lowest score as a referent. It might be adequate to take into account the opinion of several judges, as different listeners may focus their attention on different problems.

The problem of pronunciation quality evaluation in CAPT for 2LL has been addressed by several authors by making extensive use of scores obtained with subword models. Those scores are usually combined with flat weights to produce a sentence-based objective evaluation. Pronunciation quality scores have usually been based on duration, syllabic-timing, and hidden Markov model (HMM) log-likelihoods [1], [2]. Initially, those features or confidence metrics attempted to compare the observed signal with native and nonnative models by employing the forced Viterbi algorithm [2]. Also shown in [2], the phoneme log-posterior score, which is based on the Bayes classification rule for a single feature, leads to a higher correlation between subjective and objective evaluations than the ordinary features or confidence metrics by themselves. In addition, the use of nonnative *a priori* acoustic information can be considered as an important source of information to improve the evaluation accuracy by increasing the discriminability between correct and incorrect pronunciations [3].

Combining several features or confidence metrics to evaluate pronunciation quality has also been used in recent papers. In [4], a statistical combination of some measures was implemented to make a more robust automatic score pronunciation assessment. Some of the confidence measures that are usually employed in CAPT are log-likelihood ratio between native English and nonnative English HMMs given a spoken sentence, log-likelihood ratio between native English and nonnative English HMMs at the phoneme level, phoneme recognition rate, and word recognition rate. In [5], a Bayesian network structure with four metrics was proposed to take into consideration the possible pronunciation errors, to jointly evaluate pronunciation quality and reading skills. In [6], ASR technology was employed to model nonnative speaker pronunciation mistakes as phonetic variants in automatically generated competitive lexicons without the use of a detailed study of common pronunciation mistakes in an isolated-word pronunciation assessment.

Surprisingly, the problem of assessing the pronunciation quality in sentences has not been addressed as a different task and has not been modeled as the combination of the pronunciation quality of words. In [2], [7]-[12], several features, such as average phone confidence, time normalized phone confidence, and fricative confidence were combined to evaluate the pronunciation quality of sentences. The pronunciation quality of sentences has not been modeled explicitly as the combination of the pronunciation quality of words. For instance, in [8], [13], sentences were considered as scoring units. However, it is very noticeable that one 2LL teacher can score an utterance differently than another, depending on the adopted criterion, based on the pronunciation of each word. For instance, the whole sentence subjective evaluation may be determined by the worst pronounced word.

The contribution of this paper concerns the following: a) the definition of the sentence-based pronunciation evaluation task in the context of subjective criteria; b) modeling subjective criteria for the pronunciation evaluation of sentences; c) an analysis of the subjective pronunciation evaluation of sentences, based on the combination of word-based assessment; d) the generation of objective sentence pronunciation scores as the combination of word-based scores; e) a method to assess the pronunciation quality of sentences in 2LL without the need of *a priori* studies of common mistakes; and f) the use of a class-based language model to automatically incorporate a competitive lexicon and students' mother and target language phonetic rules to represent incorrect and correct pronunciations. The results presented here suggest that subjective-objective score correlation for sentences as high as 0.5 (interannotator correlation is equal to 0.65), with five levels of pronunciation quality, can be achieved. As mentioned above, the proposed method does not require any analysis of common mistakes. Controlled environments are not required to obtain the results reported herein, which are achieved using inexpensive desktop microphones. The observed performance is comparable to performances achieved using methods reliant upon *a priori* studies of pronunciation errors, which in turn also require a complete definition of the target utterances. Consequently, the integration of new target sentences with the ASR-based pronunciation quality evaluation technology is more efficient and requires less human assistance in the scheme presented here.

II. Automatic Pronunciation Assessment of Sentences with ASR

As discussed below, assessing the pronunciation of sentences in 2LL corresponds to a much more complex problem than the pronunciation evaluation of single words. Several criteria can be applied by a human evaluator to assess the pronunciation quality of sentences. For instance, the subjective score associated with a single word, $SubjWordScore_W$, could be defined based on the acoustic production quality of phonemes [6]. In contrast, the reference subjective score associated with a whole sentence, $SubjSentenceScore_S$, depends on at least one of the following three criteria that can be employed by the teacher: the minimum $SubjWordScore_W$ within the sentence,

the perceived average of $SubjWordScore_W$ within the sentence, and first impression. On the other hand, three possible combinations of objective word score, $ObjWordScore_W$, in the sentence can be considered to estimate the objective sentence score, $ObjSentenceScore_S$: the minimum $ObjWordScore_W$, the averaged $ObjWordScore_W$, and the mode of $ObjWordScore_W$.

1. Subjective Score Criteria in Sentences

Consider a target sentence $S_m = \{W_{m,1}, W_{m,2}, W_{m,3}, \ldots, W_{m,l}, \ldots, W_{m,L_m}\}$ composed of L_m words, where $W_{m,l}$ denotes the l-th word. As mentioned above, $SubjSentenceScore_{S_m}$ could be the result of one of the following criteria applied by the target language human expert.

- **Subjective Criterion 1 (SubjCrit1)**: The subjective pronunciation score of target sentence S_m corresponds to the lowest subjective score associated with one of the words $W_{m,l}$, where $1 \le l \le L_m$:

$$SubjSentenceScore_{S_m} = \min_{1 \le l \le L_m} \left\{ SubjWordScore_{W_{m,l}} \right\}. \quad (1)$$

- **Subjective Criterion 2 (SubjCrit2)**: The subjective pronunciation score of target sentence S_m corresponds to the average of the subjective scores associated with words $W_{m,l}$, where $1 \le l \le L_m$:

$$SubjSentenceScore_{S_m} = \frac{1}{L_m} \cdot \sum_{l=1}^{L_m} SubjWordScore_{W_{m,l}}. \quad (2)$$

- **Subjective Criterion 3 (SubjCrit3)**: The subjective pronunciation score of target sentence S_m is determined by the first impression without an explicit analysis on each pronounced word. Basically, in this case, the subjective evaluation is given after having heard a recorded utterance only once.

2. Objective Sentence Scores as a Combination of Word-Based Objective Scores

If each word $W_{m,l}$ in sentence S_m, where $1 \le l \le L_m$, is associated with an objective score $ObjWordScore_{W_{m,l}}$, then objective sentence score $ObjSentenceScore_{S_m}$ can be estimated by employing one of the following metric combinations.

- **Objective Metric Combination 1 (ObjMetrComb1):** The objective pronunciation score of target sentence S_m corresponds to the lowest objective score associated with one of the words $W_{m,l}$, where $1 \le l \le L_m$:

$$ObjSentenceScore_{S_m} = \min_{1 \le l \le L_m} \left\{ ObjWordScore_{W_{m,l}} \right\}. \quad (3)$$

- **Objective Metric Combination 2 (ObjMetrComb2):** The objective pronunciation score of target sentence S_m corresponds to the average word-based objective score:

$$ObjSentenceScore_{S_m} = \frac{1}{L_m} \cdot \sum_{l=1}^{L_m} ObjWordScore_{W_{m,l}}. \quad (4)$$

- **Objective Metric Combination 3 (ObjMetrComb3):** The objective pronunciation score of target sentence S_m is equal to the statistic mode or most frequent word score within S_m:

$$ObjSentenceScore_{S_m} = \max_{1 \le l \le L_m} \left\{ Frequency\left[ObjWordScore_{W_{m,l}} \right] \right\}, \quad (5)$$

where $Frequency\left[ObjWordScore_{W_{m,l}} \right]$ indicates how many times word score $ObjWordScore_{W_{m,l}}$ appears in S_m.

3. Subjective and Objective Score Correlation in Sentences

At this point, the correspondence between the subjective score criteria in subsection II.1 and the objective score combinations in subsection II.2 seems straightforward in the following two cases: SubjCrit1 and ObjMetrComb1; SubjCrit2 and ObjMetrComb2. However, SubjCrit3, considered the first impression, is much more difficult to define. A possibility is to model SubjCrit3 with the mode of the objective metrics within the target utterance (that is, ObjMetrComb3). In this sense, the best objective score combination matching is evaluated for each subjective criterion. In this paper, the accuracy of objective metric combinations is estimated by means of the correlation between the subjective scores provided by human experts and the ASR-based objective metrics.

III. ASR-Based Objective Metric with Competitive Vocabulary and Class-Based Language Model

In this paper, $ObjSentenceScore_{S_m}$ is estimated as a combination of $ObjWordScore_{W_{m,l}}$, which is in turn obtained by a continuous speech recognition system with a class-based language model [14], [15]. Given target sentence S_m as defined above, competitive class $Class_m = \{Class_{m,1}, Class_{m,2}, \ldots, Class_{m,l}, \ldots, Class_{m,L_m}\}$ is defined, where $Class_{m,l}$ can be composed of word $W_{m,l}$ according to the target pronunciation, a competitive lexicon, and phonetic variants of $W_{m,l}$. Both the competitive lexicon and phonetic variants require no *a priori* analysis of common mistakes and are automatically generated. The continuous speech recognition system is employed to make the target pronunciation of sentence S_m compete with the pronunciation of sentences composed of similar words and phonetic variants of target words. To do so, a class-based

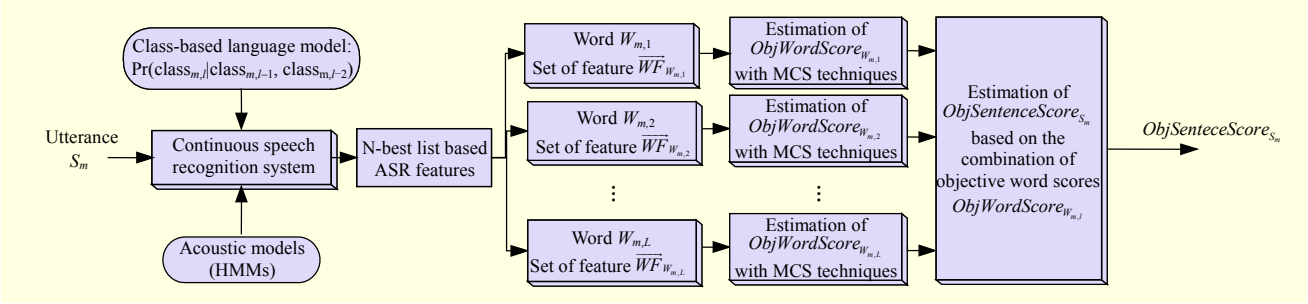

Fig. 1. Block diagram of proposed method for pronunciation evaluation of sentences.

trigram language model is generated for each sentence S_m by estimating $Pr(Class_{m,l}|Class_{m,l-1},Class_{m,l-2})$. Figure 1 shows the block diagram of the proposed scheme to assess pronunciation quality of sentences based on continuous speech recognition. Per each target word, N-best list analysis resulting from Viterbi decoding delivers a set of J word features, $\overrightarrow{WF}_{W_{m,l}}=\left[WF^1_{W_{m,l}},WF^2_{W_{m,l}},...,WF^j_{W_{m,l}},...,WF^J_{W_{m,l}}\right]$, which are detailed later. By making use of the Bayes decision rule, each word feature $WF^j_{W_{m,l}}$ is mapped to an objective $ObjWordScore^j_{W_{m,l}}$. The word features are then combined by employing multiclassifier fusion techniques to obtain the objective pronunciation metric associated with word $W_{m,l}$, $ObjWordScore_{W_{m,l}}$. Finally, the objective pronunciation score that corresponds to sentence S_m is obtained by combining $ObjWordScore_{W_{m,l}}$, where $1 \le l \le L_m$, according to subsection II.2.

1. Automatic Generation of Competitive Class

Each word within the target sentence generates a class composed of a) target word $W_{m,l}$ with the correct pronunciation, b) a competitive lexicon similar to the target word with the correct pronunciation, and c) phonetic variants of the target word according to the target or the student's native language. As a result, class $Class_{m,l}$ can be represented as

$$Class_{m,l}=\left\{W_{m,l},CL_{m,l},PV_{m,l}\right\}, \quad (6)$$

where $CL_{m,l}=\left\{CL^1_{m,l},CL^2_{m,l},CL^3_{m,l},...,CL^k_{m,l},...,CL^{K_{m,l}}_{m,l}\right\}$ denotes the competitive lexicon composed of words $CL^k_{m,l}$, where $1 \le k \le K_{m,l}$ and $K_{m,l}$ is the number of words in $CL_{m,l}$ and $PV_{m,l}$ denotes the phonetic variants of the target word according to the target and the student's native language within $Class_{m,l}$. No previous analysis based on errors made by students is required to achieve an efficient integration of didactic material to the ASR technology without human assistance. Observe that the definition and automatic generation of $Class_{m,l}$ attempt to find a tradeoff between the accuracy of the pronunciation assessment and the limitation of the ASR technology: the higher the number of competing words and phonetic variants, the more difficult the recognition task itself. It is worth mentioning that the competitive lexicon and the phonetic variants are generated by employing an acoustic-phonetic criterion only. The motivation is to optimize the performance of ASR technology, which in turn shows an inherent accuracy and limitation in pronunciation quality evaluation tasks. Consequently, the syntactic structure of the target sentence is not taken into consideration in the application addressed here. The automatic generation of $CL_{m,l}$ and $PV_{m,l}$ is described as follows.

A. Automatic Generation of Competitive Lexicon

Competitive vocabulary $CL_{m,l}$ helps to force the simultaneous competition of the correct and incorrect pronunciation and is crucial to make ASR technology successful in CAPT. This paper employs the same approach proposed in [6]. First, the Kullback-Leibler (K-L) distance defined in [17] between target word $W_{m,l}$ and the words from a lexicon representative of the target language is estimated. It is worth highlighting that the K-L distance between the target word and the words from a lexicon is normalized with respect to the phoneme alignment length [6]. Second, the lexicon whose distance to the target word is within an interval defined by a minimum, $D^{CL}_{\min}$, and a maximum, $D^{CL}_{\max}$, threshold is sorted with respect to the distance to the target word and uniformly sampled to reduce the number of selected words to the maximum number of competitive words (MNCW). Parameters $\left[D^{CL}_{\min};D^{CL}_{\max}\right]$ define a tradeoff between the discrimination ability resulting from the distance between the competitive lexicon and the target word and the accuracy of the speech recognition technology.

B. Automatic Generation of Phonetic Variants of Target Words Based on Mother and Target Languages

To improve the accuracy of the pronunciation quality evaluation, variants of the phonetic realization of target word $W_{m,l}$ according to the target and the student's native language (in this case, Spanish), $PV_{m,l} \subset \left\{PV^1_{m,l},PV^2_{m,l},PV^3_{m,l},PV^4_{m,l}\right\}$,

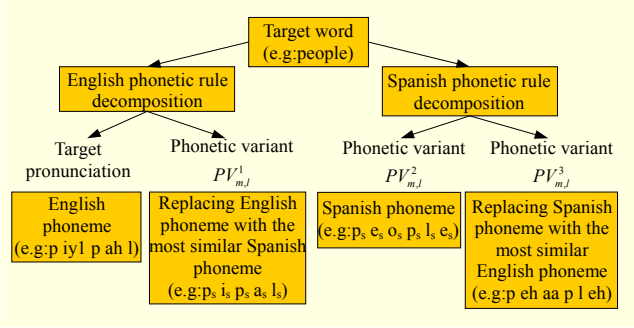

Fig. 2. Generation of phonetic variants.

are included in the competitive class $Class_{m,l}$. This strategy attempts to incorporate information about the user's native language without implementing a detailed study of pronunciation mistakes made by students. The phonetic variants are generated as follows.

As shown in Fig. 2, target word $W_{m,l}$ can be decomposed according to English or the phonetic rules of the student's native language (in this case, Spanish). In the case of English phonetic decomposition, there are two possibilities: using English phonemes or replacing English phonemes with the most similar phonemes in the student's native language according to the data listed in Table 1. In the case of decomposition according to the phonetic rules of the student's native language, there are also two possibilities: employing Spanish phonemes or replacing Spanish phonemes with the most similar phonetic units in English according to the data listed in Table 2. It is worth highlighting that Tables 1 and 2 are generated by an expert in the English language and phonetics. As a result, the phonetic variant component $PV_{m,l}$ in $Class_{m,l}$ according to (6) is defined as follows.

- $PV^1_{m,l}$ is the decomposition of target word $W_{m,l}$ according to English language phonetic rules by replacing English phonemes with those of the student's language that are most similar according to the data listed in Table 1.
- $PV^2_{m,l}$ is the decomposition of target word $W_{m,l}$ according to the phonetic rules and phonemes of the student's language.
- $PV^3_{m,l}$ is the decomposition of target word $W_{m,l}$ according to phonetic rules of the student's language by replacing phonemes with those of English that are the most similar according to the data in Table 2.
- $PV^4_{m,l}$ includes $PV^1_{m,l}$, $PV^2_{m,l}$, and $PV^3_{m,l}$ simultaneously in $Class_{m,l}$.

Then, $PV^i_{m,l}$, where $1 \le i \le 4$, is included in $PV_{m,l}$ if the K-L distance between $PV^i_{m,l}$ and target word $W_{m,l}$ is greater than or equal to $D^{PV}_{\min}$. Threshold $D^{PV}_{\min}$ defines a tradeoff between the discrimination ability resulting from the distance between the phonetic variants and the target word and the accuracy of the speech recognition technology.

Table 1. English phonemes are replaced with most similar Spanish phonemes to generate phonetic variant $PV^1_{m,l}$.

English	Spanish	English	Spanish	English	Spanish
Ah, Ae,	a_s	Hh	j_s	Zh, Jh, Y	y_s
Aa, Ao	o_s	K	k_s	T	t_s
B, V	b_s	L	l_s	Uh, Uw	u_s
Ch, Sh	ch_s	M	m_s	Oy	o_s i_s
D, Dh	d_s	N	n_s	Aw	a_s u_s
Eh, Er	e_s	Ow	o_s u_s	W	g_s u_s
F	f_s	P	p_s	Ng	n_s g_s
G	g_s	R	r_s	Ay	a_s i_s
Ih, Iy	i_s	S, Z, Th	s_s	Ey	e_s i_s

Table 2. Spanish phonemes are replaced with most similar English phonemes to generate phonetic variant $PV^3_{m,l}$.

Spanish	English	Spanish	English	Spanish	English
a_s	Ah	i_s	Ih	p_s	P
b_s	B	j_s	Hh	r_s	R
ch_s	Ch	k_s	K	s_s	S
d_s	D	l_s	L	t_s	T
e_s	Eh	m_s	M	u_s	Uh
f_s	F	n_s	N	y_s	Y
g_s	G	o_s	Aa		

2. Class-Based Language Model in ASR

Continuous speech recognition is run by using a class-based language model [14], [15]. As mentioned above, competitive class $Class_m$ is generated from sentence S_m. Then, trigrams $Pr(Class_{m,p}|Class_{m,q},Class_{m,r})$ are defined as follows.

$$Pr\left(Class_{m,p}\,|\,Class_{m,q},Class_{m,r}\right)=\begin{cases}1 & \text{if } q=p-1 \wedge r=q-1 \wedge p\ge 3,\\ 0 & \text{otherwise.}\end{cases} \tag{7}$$

When p=2, $Pr(Class_{m,p}|Class_{m,q},Class_{m,r})$ is replaced with bigram $Pr(Class_{m,p}|Class_{m,q})$.

$$Pr\left(Class_{m,p}\,|\,Class_{m,q}\right)=\begin{cases}1 & \text{if } q=p-1 \text{ and } p=2,\\ 0 & \text{otherwise.}\end{cases} \tag{8}$$

The class-based language model attempts to identify mispronounced words within the target sentence.

3. Word-Based N-Best List Feature Extraction

Given a target utterance, ASR with a class-based language

model enables the efficient extraction of several features per word. In this paper, four confidence measures delivered by the ASR procedure are employed: N-best position, recognition flag, word density, and logarithmic word density.

The position in the N-best list of word $W_{m,l}$ in target sentence S_m, $POS_{m,l}$, corresponds to the index of the most likely hypothesis in which $W_{m,l}$ is recognized:

$$POS_{m,l} = argmax_r \left\{ \left[Q(h_r) \right] \middle| r \in E(W_{m,l}, H) \right\}, \quad (9)$$

where $Q(h_r) = P(h_r)^{\gamma} \cdot P(O/h_r)$, h_r is the r-th hypothesis in the N-best Viterbi list, $Q(h_r)$ is the likelihood score given by the Viterbi search, $P(h_r)$ is the language model probability of h_r, $P(O/h_r)$ is the observation probability of h_r, γ is the acoustic model scaling factor, $E(W_{m,l}, H)$ corresponds to the indices of the hypotheses in which word $W_{m,l}$ is contained, and, finally, H denotes all the N-best alignments or hypotheses obtained from Viterbi decoding.

The recognition flag binary confidence measure associated with word $W_{m,l}$ in target sentence S_m, denoted by $REC_{m,l}$, is defined as

$$REC_{m,l} = \begin{cases} 1 & \text{if } W_{m,l} \in h_1, \\ 0 & \text{if } W_{m,l} \notin h_1, \end{cases} \quad (10)$$

where h_1 is the first hypothesis in the N-best Viterbi list. The word density confidence measure of word $W_{m,l}$ in target sentence S_m, $WDCM_{m,l}$, is defined as [16]

$$WDCM_{m,l} = \frac{\sum_{r \in E(W_{m,l},H)} Q(h_r)}{\sum_{l=1}^{N} Q(h_l)}, \quad (11)$$

where $Q(h_r)$ and $E(W_{m,l}, H)$ are defined as above. The logarithmic word density confidence measure of target word $W_{m,l}$, $LogWDCM_{m,l}$, is defined as

$$LogWDCM_{m,l} = \frac{\sum_{r \in E(W_{m,l},H)} \log(Q(h_r))}{\sum_{l=1}^{N} \log(Q(h_l))}, \quad (12)$$

where $Q(h_r)$ and $E(W_{m,l}, H)$ are defined as above.

4. Word-Based Objective Pronunciation Score Estimation

As in [6], the word-based objective pronunciation score, $ObjWordScore_{W_{m,l}}$, is estimated by employing multiclassifier system (MCS) techniques, as shown in Fig. 1. As described above, four word features or confidence metrics are evaluated per word in the target sentences: $POS_{m,l}$, $REC_{m,l}$, $WDCM_{m,l}$, and $LogWDCM_{m,l}$. The problem of word-based pronunciation quality evaluation is modeled as a mapping between confidence metrics and score $ObjWordScore_{W_{m,l}}$, which emulates the opinion given by a human instructor, $SubjWordScore_{W_{m,l}}$. Suppose that subjective score $SubjWordScore_{W_{m,l}}$ is quantized in V levels (in this paper, V=5). Consequently, every confidence metric could be assumed to be a score delivered by a given classifier, and every subjective score level would be a class. Consider that O is the sequence of observation vectors corresponding to target sentence S_m uttered by a student. By using the Bayes rule, $ObjWordScore_{W_{m,l}}$ can be estimated as

$$ObjWordScore_{W_{m,l}}(O) = \underset{C_v}{argmax} \left\{ \frac{P(\overrightarrow{WF}_{W_{m,l}}(O)/C_v)P(C_v)}{P(\overrightarrow{WF}_{W_{m,l}}(O))} \right\}, \quad (13)$$

where $ObjWordScore_{W_{m,l}}(O)$ is the final decision for $W_{m,l}$ and corresponds to signal O and $1 \le v \le V$. Theoretically, the classification error is optimally minimized by (13). $P(C_v)$ is assumed uniformly distributed and equal to $1/V$. However, the *a priori* multivariable probability density functions (PDFs) $P(\overrightarrow{WF}_{W_{m,l}}(O)/C_v)$ and $P(\overrightarrow{WF}_{W_{m,l}}(O))$ may require an unmanageable amount of training data to be estimated reliably [18]. As a consequence, the problem is substantially simplified if maximization in (13) can be expressed in terms of computations performed by individual classifiers. The classical techniques to simplify the Bayesian fusion [18]-[20] are product rule, max rule, min rule, mean rule, and majority vote rule (MVR). Among the several MCS combination rules in the literature, mean rule and MVR are the most frequently employed approximations to simplify the Bayesian fusion [21], [22]. Product rule corresponds to the optimal Bayesian fusion if the classifiers are statistically independent. MVR allows combining local decisions of individual classifiers. The mean rule is defined as

$$ObjWordScore_{W_{m,l}}(O) = argmax_{C_v} \left\{ \frac{1}{J} \sum_{j=1}^{J} P(C_v / WF^j_{W_{m,l}}(O)) \right\}$$
$$= argmax_{C_v} \left\{ \frac{1}{J} \sum_{j=1}^{J} \frac{P(WF^j_{W_{m,l}}(O)/C_v)P(C_v)}{P(WF^j_{W_{m,l}}(O))} \right\}. \quad (14)$$

As mentioned above, $1 \le v \le V$, and the total number of possible levels of pronunciation quality is V, and the total number of word features or confidence metrics is J.

MVR is a straightforward scheme to combine the output of individual classifiers [21]. Given a set of individual classifier decisions, the final decision will be the class that receives the largest number of votes as the consensus.

IV. Experiments

The native American English acoustic models were trained

with the CSR-I WSJ0[2)] corpus [23]. In CSR-I WSJ0, speech data was recorded with a high-quality microphone and the sample rate was equal to 16 kHz. All the 20,055 training utterances in CSR-I WSJ0 are used to train English CDHMMs (approximately 40 hours of speech). Also, LATINO 40 [24] is employed to train the Spanish phonetic units used to generate the phonetic variants according to part B of subsection III.1. This database is composed of continuous speech from 40 Latin American native speakers, with each speaker reading 125 sentences from newspapers in Spanish (approximately 2.5 hours of speech). The training utterances were 4,500 WAV PCM sentences provided by 36 speakers and context-dependent phoneme HMMs were employed. The vocabulary is composed of almost 6,000 words. The CSR-I WSJ0 and Latino-40 databases are selected because they were recorded by native speakers in a controlled environment with high quality microphones. Also, they correspond to medium vocabulary, which in turn guarantees that they are representative of the phonetic variability in English and Spanish. The difference in size is compensated by the number of triphonemes to train: 12,491 in CSR-I WSJ0 and 2,447 in Latino-40. Also, recognition experiments with Latino-40 show that the database size is enough to lead to state-of-the-art word error rates. Thirty-three MFCC parameters per frame are computed: the frame energy plus ten static coefficients and their first and second time derivatives. The number of static features is chosen to optimize the recognition accuracy. Observe that the accuracy of the method presented here is extremely dependable for speech recognition technology. Cepstral mean normalization (CMN) is also employed. Each monophone and triphoneme is modeled with a three-state left-to-right topology without a skip-state transition, with eight multivariate Gaussian densities per state with diagonal covariance matrices. The language model is estimated according to subsection III.2. As explained above, a competitive class, as defined in (6), is generated for each target word within a given target sentence. The competitive lexicon for each target word is chosen from the vocabulary that composes the CSR-I WSJ0 corpus, as in [6]. The phonetic variants of each target word are generated according to part B of subsection III.1 and Fig. 2. Four confidence measures delivered by the ASR procedure are employed: $POS_{m,l}$, $REC_{m,l}$, $WDCM_{m,l}$, and $LogWDCM_{m,l}$.

The data base is composed of 423 utterances from 43 speakers with different levels of English proficiency and is recorded using inexpensive desktop microphones. The sentences are extracted from a web-based 2LL system and are selected by an expert in the English language and phonetics to achieve a phonetically balanced evaluation data set. Examples of the sentences contained are "It was nice to see my relatives" and "I missed my Kiwi family." The database is arbitrarily divided into subset 1 (212 utterances) and subset 2 (211 utterances). The experiments are performed by using subset 1 and subset 2 as training and testing data, respectively, and vice-versa. Consequently, each experiment employs all 423 utterances. The training data is employed to estimate the *a priori* PDFs in (14), which correspond to nonparametric distributions (that is, histograms).

The subjective scores are determined with seven experts in the English language. The pronunciation quality of each word within target utterance l, $SubjWordScore_{W_{m,l}}$, is evaluated by two experts in the English language, as in [6]. If the evaluations diverge from one another, the opinion of a third expert is taken into consideration. The scoring is done using a 1 to 5 scale: score 5 corresponds to the correct pronunciation of the target word, and score 1 denotes the worst possible pronunciation, generally the result of the application of Spanish pronunciation rules. Then, subjective scores SubjCrit1 and SubjCrit2 in each sentence are estimated with $SubjWordScore_{W_{m,l}}$, as explained in subsection II.1. In contrast, subjective score SubjCrit3, which corresponds to the first impression, is determined by asking the English language experts to give their opinions after listening to each sentence only once. The interannotator correlation is equal to 0.65.

V. Discussion

Tables 3 through 5 show the subjective-objective score correlation versus threshold $D_{\min}^{PV}$, defined in part B of subsection III.1. As explained above, $D_{\min}^{PV}$ defines a minimum distance threshold between the target pronunciation and the phonetic variance due to the fact that the ASR technology accuracy imposes a higher bound for discrimination ability. Competitive class $Class_{m,l}$ in (6) is composed of the target pronunciation of $W_{m,l}$, competitive lexicon $CL_{m,l}$, and phonetic variant $PV_{m,l}$. The phonetic variant included in $Class_{m,l}$ is the one that gives the highest subjective-objective score correlation and is chosen among $PV_{m,l}^1$, $PV_{m,l}^2$, $PV_{m,l}^3$, and $PV_{m,l}^4$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ are estimated according to SubjCrit2 and ObjMetrComb2, SubjCrit1 and ObjMetrComb1, and SubjCrit3 and ObjMetrComb3.

As shown in Table 3, $CL_{m,l}$ is generated with *MNCW* equal to 5, $\left[D_{\min}^{CL}=8;\ D_{\max}^{CL}=25\right]$, and $PV_{m,l}=\left\{PV_{m,l}^3\right\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ are estimated according to SubjCrit2 in (2) and ObjMetrComb2 in (4), respectively. According to Table 3, the maximum subjective-

2) CSR-I (WSJ0) Sennheiser, Publisher by LDC, ISBN: 1-58563-006-3

Table 3. Subjective-objective score correlation vs. threshold $D_{\min}^{PV}$. $CL_{m,l}$ is generated with $MNCW$=5, $[D_{\min}^{CL} = 8; D_{\max}^{CL}=25]$, and $PV_{m,l}=\{PV_{m,l}^{3}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ are estimated according to SubjCrit2 and ObjMetrComb2.

$D_{\min}^{PV}$	MCS	
	Mean rule	MVR
0	0.45 (0.1736)	0.48 (0.2207)
1	0.5	0.52
2	0.47 (0.2843)	0.49 (0.281)
3	0.46 (0.2266)	0.47 (0.1685)
4	0.43 (0.0968)	0.45 (0.0918)

Table 4. Subjective-objective score correlation vs. threshold $D_{\min}^{PV}$. $CL_{m,l}$ is generated with $MNCW$=10, $[D_{\min}^{CL}=8; D_{\max}^{CL}=25]$ and $PV_{m,l}=\{PV_{m,l}^{1}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ were estimated according to SubjCrit1 and ObjMetrComb1.

$D_{\min}^{PV}$	MCS	
	Mean rule	MVR
0	0.5 (0.2776)	0.49 (0.2177)
2.5	0.49 (0.2177)	0.5 (0.2776)
5	0.53	0.53
7.5	0.52 (0.4207)	0.53 (0.5)
10	0.51 (0.3446)	0.5 (0.2776)

objective score correlation is achieved when $D_{\min}^{PV}=1$ with the MCS mean and MVR. The statistical significances of the difference with respect to the highest subjective-objective score correlation are presented in parentheses. $D_{\min}^{PV}$ can introduce increases of 11.1% and 8.3% in the subjective-objective score correlation with MCS mean rule and MVR, respectively, when results are compared with the case in which no threshold is applied, that is, $D_{\min}^{PV}=0$.

As shown in Table 4, $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ are estimated according to SubjCrit1 in (1) and ObjMetrComb1 in (3). Competitive lexicon $CL_{m,l}$ in (6) is obtained with $MNCW$ equal to 10, $[D_{\min}^{CL}=8; D_{\max}^{CL}=25]$, as explained in part A of subsection III.1, and $PV_{m,l}=\{PV_{m,l}^{1}\}$. The statistical significances of the differences with respect to the highest subjective-objective score correlation are presented in parentheses. Table 4 shows that $D_{\min}^{PV}$ can introduce increases of 6.0% and 6.1% in the subjective-objective score correlation with MCS mean rule and MVR, respectively, when results are compared with the case in which no threshold is applied, that is, $D_{\min}^{PV}=0$.

Table 5. Subjective-objective score correlation vs. threshold $D_{\min}^{PV}$. $CL_{m,l}$ is generated with $MNCW$=15, $[D_{\min}^{CL}=8; D_{\max}^{CL}=25]$, and $PV_{m,l}=\{PV_{m,l}^{4}\}$. $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ are estimated according to SubjCrit3 and ObjMetrComb3.

$D_{\min}^{PV}$	MCS	
	Mean rule	MVR
0	0.52	0.53
1	0.46 (0.1251)	0.48 (0.166)
2	0.43 (0.0455)	0.43 (0.0294)
3	0.4 (0.0136)	0.42 (0.0197)
4	0.4 (0.0136)	0.41 (0.0125)

As shown in Table 5, $SubjSentenceScore_{S_m}$ and $ObjSentenceScore_{S_m}$ are estimated according to SubjCrit3 as defined in subsection II.1 and ObjMetrComb3 as defined in (5). Competitive lexicon $CL_{m,l}$ in (6) is obtained with $MNCW$ equal to 10 and $\left[D_{\min}^{CL}=8;\ D_{\max}^{CL}=25\right]$, as explained in part A of subsection III.1. Phonetic variant $PV_{m,l}$ is made equal to $PV_{m,l}^{4}$. The statistical significances of the differences with respect to the highest subjective-objective score correlation are presented in parentheses. As shown in Table 5, the maximum subjective-objective score correlation is achieved when $D_{min}^{PV}=0$ with the MCS mean rule and MVR.

In Table 6, each subjective criterion is modeled with ObjMetrComb1, ObjMetrComb2, and ObjMetrComb3 by using MCS MVR. As can be seen in Table 6, SubjCrit1 and SubjCrit2 can be modeled more accurately with ObjMetrComb1 and ObjMetrComb2, respectively. The subjective-objective correlation obtained by employing SubjCrit3/ObjMetrComb3 is high and comparable with those achieved with SubjCrit1/ObjMetrComb1 and SubjCrit2/ ObjMetrComb2. However, the highest subjective-objective score correlation in the SubjCrit3 column is achieved with ObjMetric2 instead of ObjMetric3, as shown in Table 6. This result suggests that the first impression criterion may employ an averaging procedure of word-based pronunciation assessment. This result could also be due to the fact that the average word-based objective score is a more robust estimation than the statistic mode or most frequent word score within the target sentence. The statistical significances of the differences with respect to the highest subjective-objective score correlation in each column are presented in parentheses. A similar result is obtained by replacing MVR with the mean rule as the MCS method.

As mentioned above, the competitive lexicon and the phonetic variants are generated by employing an acoustic-phonetic criterion only. Consequently, the syntactic structure of

Table 6. Subjective-objective score correlation with MCS MVR. Each subjective criterion is modeled with ObjMetrComb1, ObjMetrComb2, and ObjMetrComb3.

MCS MVR	SubjCrit 1	SubjCrit 2	SubjCrit 3
ObjMetrComb 1	0.53	0.29 (0)	0.29 (0)
ObjMetrComb 2	0.51 (0.3483)	0.52	0.57
ObjMetrComb 3	0.27 (0)	0.44 (0.0708)	0.53 (0.209)

Table 7. Subjective-objective score correlation published elsewhere with similar tasks.

Ref.	Average obj-subj correlation (no. of score levels)	Interannotator correlation
[3]	0.453 (5 levels)	0.65
[4]	0.58 (5 levels)	0.65
[12]	0.4913 (5 levels)	0.61
[13]	0.521 (5 levels)	0.65
[14]	0.642(5 levels)	0.65
[15]	0.415 (10 levels)	0.593

the target sentence is not taken into consideration in the task addressed here.

Finally, Table 7 shows results achieved by other authors with similar tasks. When compared with the methods in Table 7, the proposed technique leads to similar results but shows the following advantages: It offers the possibility to define more than one criterion to combine word scores and emulate subjective criteria; It does not require *a priori* studies of common mistakes; It does not require controlled environments nor high quality microphones.

VI. Conclusion

This paper proposed that the problem of sentence-based pronunciation evaluation tasks should be defined in the context of subjective criteria. Three subjective criteria (that is, the minimum subjective word score, the mean subjective word score, and first impression) were proposed for the pronunciation evaluation of sentences and modeled based on the combination of word-based assessment. Then, the subjective criteria were approximated with objective sentence pronunciation scores obtained with the combination of word-based metrics. As the proposed method does not need *a priori* studies of common mistakes, class-based language models were used to incorporate the student's native and target language phonetic rules to represent incorrect and correct pronunciations. Incorrect pronunciations were automatically generated by incorporating a competitive lexicon and applying students' native and target language phonetic rules, applicable to any 2LL context.

The results presented here suggest that subjective-objective sentence score correlations greater than or equal to 0.5 can be achieved when the proposed sentence-based pronunciation criteria are approximated with the combination of word-based scores. By considering that the interannotator correlation is 0.65, the achieved subjective-objective sentence score correlations can be interpreted as very positive results. Particularly, it is worth emphasizing that the minimum subjective word score, the mean subjective word score, and the first impression can effectively be emulated with the lowest word objective score within the target sentence, the average word-based objective score, and the statistic mode or most frequent word score in the utterance, respectively. This result is especially interesting to emulate more than one subjective criterion in pronunciation evaluation.

The subjective-objective sentence score correlations achieved here are very comparable with those published elsewhere with *a priori* studies of pronunciation errors. As a consequence, the integration of new target sentences with the ASR-based pronunciation quality evaluation technology is more efficient and requires less human assistance with the approach proposed in this paper. Improving the discrimination ability provided by ASR-based technology between correct and incorrect pronunciations, proposing more accurate models for the subjective first impression criteria, and proposing new subjective criterion (for example, semantic) are proposed for future research.

References

[1] L. Neumeyer et al., "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech," *Proc. ICSLP*, 1996, pp. 1457-1460.

[2] H. Franco et al., "Automatic Pronunciation Scoring for Language Instruction," *ICASSP*, vol. 2, 1997, pp. 1471-1474.

[3] A. Neri, C. Cucchiarini, and W. Strik, "Automatic Speech Recognition for Second Language Learning: How and Why It Actually Works," *Proc. 15th Int. Congress Phonetic Sci.*, Barcelona, Spain, 2003, pp. 1157-1160.

[4] S. Nakagawa and K. Ohta, "A Statistical Method of Evaluating Pronunciation Proficiency for Presentation in English," *Proc. InterSpeech*, Antwerp, Belgium, Aug. 2007.

[5] J. Tepperman et al., "A Bayesian Network Classifier for Word-Level Reading Assessment," *Proc. InterSpeech*, Antwerp, Belgium, Aug. 2007.

[6] C. Molina et al., "ASR Based Pronunciation Evaluation with Automatically Generated Competing Vocabulary and Classifier Fusion," *Speech Commun.*, vol. 51, no. 6, June 2009, pp. 485-498.

[7] O. Deshmukh, S. Joshi, and A. Verma, "Automatic Pronunciation Evaluation and Classification," *INTERSPEECH*, 2008, pp. 1721-1724.

[8] T. Cincarek et al., "Automatic Pronunciation Scoring of Words and Sentences Independent from the Non-native's First Language," *Computer Speech Language*, vol. 23, no. 1, Jan. 2009, pp. 65-88.

[9] S. Xu et al., "Automatic Pronunciation Evaluation Based on Feature Extraction and Combination," *Proc. 3rd Int. Conf. Innovative Computing Inf. Control*, 2008, pp. 172-176.

[10] N. Moustroufas and V. Digalakis, "Automatic Pronunciation Evaluation of Foreign Speakers Using Unknown Text," *Computer Speech Language*, vol. 21, no. 1, Jan. 2007, pp. 219-230.

[11] L. Neumeyer et al., "Automatic Scoring of Pronunciation Quality," *Speech Commun.*, vol. 30, no. 2-3, Feb. 2000, pp. 83-93.

[12] H. Franco et al., "Combination of Machine Scores for Automatic Grading of Pronunciation Quality," *Speech Commun.*, vol. 30, no. 2-3, Feb. 2000, pp. 121-130.

[13] S. Wei et al., "Pronunciation Space Models for Pronunciation Evaluation," *6th Int. Symp. Chinese Spoken Language Process. (ISCSLP)*, Dec. 2008, pp. 1-4.

[14] W. Ward and S. Issar, "A Class Based Language Model for Speech Recognition," *Proc. ICASSP*, 1996, pp. 416-418.

[15] J. Zhang et al., "Improvements in Audio Processing and Language Modeling in the CU Communicator," *Eurospeech*, Aalborg, Denmark, 2001.

[16] K.Y. Kwan, T. Lee, and C. Yang, "Unsupervised N-Best Based Model Adaptation Using Model-Level Confidence Measures," *Proc. ICSLP*, 2002, pp. 69-72.

[17] J. Sooful and E. Botha, "Comparison of Acoustic Distance Measures for Automatic Cross-Language Phoneme Mapping," *Proc. ICSLP*, Denver, CO, USA, 2002, pp. 521-524.

[18] J. Kittler et al., "On Combining Classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, Mar. 1998, pp. 226-239.

[19] L.I. Kuncheva, J.C. Bezdeck, and R.P.W. Duin, "Decision Templates for Multiple Classifier Fusion: An Experimental Comparison," *Pattern Recog.*, vol. 34, no. 2, 2001, pp. 299-314.

[20] L.I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, Feb. 2002, pp. 281-286.

[21] J. Kittler and F.M. Alkoot, "Sum versus Vote Fusion in Multiple Classifier Systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, issue 1, 2003, pp. 110-115.

[22] G. Fumera and F. Roli, "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, June 2005, pp. 942-956.

[23] J. Garofalo et al., *Continuous Speech Recognition (CSR-I) Wall Street Journal (WSJ0) News, Complete*, Linguistic Data Consortium, Philadelphia, PA, USA, 1993.

[24] Linguistic Data Consortium, *LATINO-40 Spanish Read News Corpus*, database, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA, 1995.

Néstor Becerra Yoma received his PhD from the University of Edinburgh, Edinburgh, Scotland, UK, and his MSc and BSc from UNICAMP (Campinas State University), Sao Paulo, Brazil, all in electrical engineering, in 1998, 1993, and 1986, respectively. Since 2000, he has been a professor in the Department of Electrical Engineering, Universidad de Chile, in Santiago, where he is currently lecturing on telecommunications and speech processing and working on robust speech recognition/speaker verification, language learning, dialogue systems, and voice over IP. In 2011, he was promoted to the full professor position. At the Universidad de Chile, he started the Speech Processing and Transmission Laboratory to carry out research on speech technology applications for the Internet and telephone line. He is the author of 24 journal papers, over 30 conference papers, and two awarded patents. Professor Becerra Yoma is one of the associate editors of *IEEE Transactions on Speech and Audio Processing*, and he is a member of the Institute of Electrical and Electronics Engineers and the International Speech Communication Association.

Leopoldo Benavides Berrios received his MSc and BSc in electrical engineering from the Universidad de Chile, Santiago, Chile, in 2011 and 2009, respectively. From 2009 to 2011, he was a research assistant at the Speech Processing and Transmission Laboratory, where he completed his MSc dissertation on computer-aided pronunciation training and language learning.

Jorge Wuth Sepúlveda received his BSc in electrical engineering from the Universidad de Chile, Santiago, Chile, in 2007. Since 2007, he has been a research associate at the Speech Processing and Transmission Laboratory, where he is currently carrying out his research on speech recognition and computer-aided pronunciation training and language learning. Mr. Wuth is the co-author of three journal articles and two conference papers. His research interests include speech recognition and web-based human machine interfaces.

Hiram Vivanco Torres was born in Santiago de Chile, October 10, 1942. He completed his undergraduate studies at the Instituto Pedagógico (Teachers Training College) of the Universidad de Chile. He completed his graduate studies at the University of Michigan, Ann Arbor, MI, USA (MA, linguistics) in 1966 and at the University of Lancaster, Lancaster, England, UK (MA, applied linguistics) in 1972. He is a full professor of phonetics and phonology in the Faculty of Philosophy and Letters of the Universidad de Chile. He has published over 70 articles in specialized national and international journals. He has been the treasurer of the Chilean Phonetics Teachers Association since 1982 and was the regional secretary for South America of the International Society of Phonetic Sciences (ISPhS) from 1987 to 1996. He was invited by the Real Academia Española and the Government of Colombia to participate in and present a paper at the IV Congreso de la Lengua Española, Cartagena de Indias, Colombia, in March 2007. He has also been invited to several conferences and workshops as a key speaker.