ON THE USE OF PHONOLOGICAL FEATURES FOR PRONUNCIATION SCORING

Frederik Stouten, Jean-Pierre Martens

ELIS, Ghent University, Sint-Pietersnieuwstraat 41, B-9000, Ghent

ABSTRACT

It is acknowledged that in many medical and educational applications there is a great need for good objective assessments of the pronunciation proficiency of a speaker, either a non-native speaker of the language or a native speaker with a certain speech handicap (e.g. a deaf or dysarthric speaker). Most pronunciation scoring software developed thus far just measures an over-all proficiency. The system proposed here envisages the computation of more detailed information on the nature of the pronunciation deficiencies. To that end, it works with a phonological representation of the speech. Described is the development of the system as well as its first encouraging assessments of non-native speakers of American English.

1. INTRODUCTION

More and more people try to learn a second language in a short time, and they use Computer Assisted Language Learning (CALL) tools as tutors. However, it is acknowledged that the effectiveness of such tools could be increased if there were better speech algorithms for detecting pronunciation deficiencies and for providing specific information about the nature of these deficiencies.

Several groups [1, 2, 3] already proposed methods for computing over-all pronunciation proficiency scores, and even measured correlations of these scores with human ratings. Although they found that the correlation on sentence-level was reasonable (=0.76), it was disappointing to find out that direct measures of pronunciation phenomena, such as log likelihoods emerging from an HMM speech recognizer, contributed less to that correlation than simple over-all acoustic features which mainly charactere the speech rate.

In the present paper an attempt is made to develop a new scoring scheme which will be based on articulatory information that can be retrieved from the acoustic signal. Such information can be beneficial for the scoring of non-native as well as pathological native speech which contains phonetic units (phonemes or subphonemic units) that are way-off the expected canonical ones. For instance, if a second language (L2) learner utters a phoneme of L2 that is atypical for his mother tongue (L1) this is likely going to result in a pronunciation that is mapped to a non-canonical point in an articulatory space. Since such a space is spanned by interpretable dimensions, the deviations may be easy to translate into meaningful feedback for the user. Our aim is to take advantage of this and to surpass the work of [4] which also employs articulatory features to distinguish between correct and incorrect pronunciations. In [5], they use a state-of-the-art speech recognizer to detect phone segments with low confidences, and a separate feedback module to provide feedback in terms of phonological features. The advantage of our approach is that it is a much more integrated appraoch. All parts work in the same feature space.

The structure of the paper is as follows. First we outline our general methodology (Section 2), then we introduce the phonological features we are going to use (Section 3) and we discuss the feature extraction as well as the automatic segmentation and labeling we derive (Section 4). In Section 5 we present some simple pronunciation proficiency scores derived from this segmentation and labeling, and we report initial results on the assessment of non-native speakers.

2. METHODOLOGY

The envisaged pronunciation proficiency scoring system starts by converting the acoustic features (ACFs) of each speech frame into a new feature vector which is believed to describe the articulatory configuration of the vocal tract during the production of that frame. However, since this feature mapping will be learned from phonological labels rather than from articulatory measurements (as used in [6]), the produced features will be called phonological features (PHFs). It is nevertheless presumed that these PHFs do describe the underlying articulatory configuration.

It is well known [7, 8] that a particular ACF vector can correspond to different articulatory configurations. This means that the proposed ACF-to-PHF mapping is actually a *one-to-many* mapping. Nevertheless, one can design non-linear functions [9] that can map the frames of the different phonemes to hardly overlapping regions in a properly defined PHF space. We have defined a slightly novel PHF set and detector architecture.

To convert the PHF vector sequence into pronunciation scores, we need either an orthographic or a phonemic transcription of the utterance under test. The PHF vectors derived from the acoustics can then be lined up with the expected sequence of phonetic units (phonemes or phoneme components) emerging from this transcription. Such a process is commonly called speech alignment or segmentation & labeling.

Our speech aligner is supplied with vectors composed of highly correlated elements. Therefore it incorporates specific techniques which are not found in a traditional HMM-based speech aligner working with MFCCs (Mel-scale Frequency Cepstral Coefficients).

Once the segmentation and labeling is accomplished it offers a framework for the computation of phoneme-specific and/or feature-specific pronunciation goodness scores.

3. PHONOLOGICAL FEATURES : DEFINITION

Although first proposed by Jakobson, Fant and Halle [10], the real power of PHFs was only demonstrated in *The Sound Pattern of English* (SPE) [11]. With 13 binary features it was possible to obtain a unique description of all the English phonemes. Since 1968, many alternative feature sets have been proposed (e.g. [9, 12]), and in [9] one proves that some of the previously proposed feature sets can be retrieved reliably from the acoustic signal by means of recurrent neural networks. Starting from this and related work we have tried to identify a feature set that meets the following two criteria: (1) on the basis of phonological knowledge, it should be easy to attribute canonical feature values to all the phonetic units, and (2) it should

be possible to extract these canonical features in a reliable way by means of an automatically trained feature mapper.

It is clear that not all phonological features are relevant for the description of all the phonetic units. Therefore it is recommended to combine features which are relevant and irrelevant at the same time into separate feature dimensions. After some experimentation we finally came to the following PHF definition involving 27 binary features encoding four feature dimensions:

- **vocal source**: voiced, unvoiced, no-activation
- **manner**: closure, vowel, fricative, burst, nasal, approximant, lateral, sil
- **place-consonant**: labial, labio-dental, dental, alveolar, post-alveolar, velar, glottal
- **vowel-features**: low, mid-low, mid-high, high, back, mid, front, retroflex, round

In this definition the vocal source is presumed to describe the frame-level presence/absence of speech excitation and the nature (voiced/ unvoiced) of that excitation. The other features of a frame are presumed to describe the properties of the phonetic unit to whose realization that frame contributes. Their detection will require inputs from broad time interval. For instance, the distinction between a *closure* and a *silence* resides in the length of the no-activation interval.

4. PHONOLOGICAL FEATURE EXTRACTION

Like in [13, 14], we propose to use a hierarchical feature extractor (Figure 1). The vocal source is retrieved directly from the ACFs, the

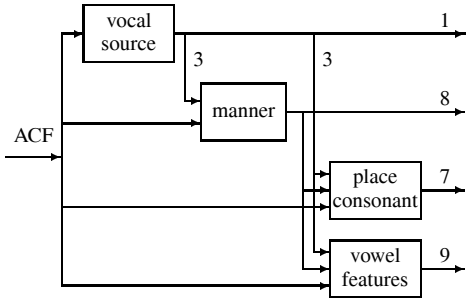

Fig. 1. The phonological feature extractor.

manner features get the vocal source output as a supplementary input and the consonant and vowel feature extraction can benefit from the manner features as well.

Since the vocal source is presumed to be a local property, its detection is based on 7 input vectors representing a window of 7 frames centered around the frame of interest. For the detection of the other features a window of 11 frames is used. However, to reduce the number of input vectors, the three most distant frames of the left and right context are represented by their mean vectors.

4.1. Training of the detectors

Each of the four detectors is implemented as a multi-layer perceptron (MLP) with one hidden layer. The training of these MLPs is performed by means of the on-line EEBP algorithm [15]. After training, the MLP outputs are supposed to represent the posterior probabilities of the PHFs. MLP weight updates are derived from deviations between the computed and the desired outputs. However, not all features are relevant for every frame, and this fact has to be accounted for. For instance, if a frame is known to contribute to the production of a consonant, the desired vowel-features are marked as *unknown* and the frame does not contribute to the training of the *vowel-features* MLP. If a frame contributes to the production of a diphthong the desired place of articulation in the vowel-features set is marked as *unknown* and the deviation between the computed and desired place outputs is assumed to be zero.

The training uses line search to adapt the learning rate and to decide whether to continue the training or not (see [15]).

4.2. Evaluation of the detectors

The feature extraction was evaluated on the manually segmented and labeled TIMIT corpus. The training was performed on 3696 utterances (420 speakers times 8 utterances), and 1344 utterances (168 speakers times 8 utterances) were available for testing. The ACFs

MLP	#inputs	#hidden	#outputs	#weights
Source	91	100	3	9503
Manner	94	250	8	25008
Place-C	101	250	7	27258
Vowel-F	101	250	9	27760

Table 1. *Nr of inputs, hidden nodes, outputs and weights of MLPs*

consisted of 12 MFCCs and a log-energy per frame (frame length of 25 ms, frame shift of 10 ms).

The number of inputs, hidden units, outputs and weights per MLP are listed in Table 4.2. Since the canonical *manner* and *place-consonant* representations show only one positive output, the corresponding MLPs can be evaluated as frame classifiers. The accuracy of the *manner* MLP was 83.9 %, that of the place-consonant MLP was 83.2 %. These figures are in line with those reported in [16, 9].

5. SEGMENTATION AND LABELING

The segmentation and labeling of an utterance is based on the alignment of the PHF vector sequence with a linguistic model derived from the orthographic or phonemic transcription of that utterance.

5.1. System architecture

The system architecture is depicted in Figure 2. The output is a sequence of starting times (t_k) and corresponding phonetic labels (p_k). If $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is the ACF sequence then the Viterbi decoder

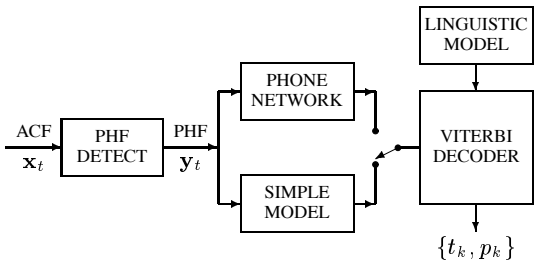

Fig. 2. System architecture

searches for the state sequence $S = \{s_1, \ldots, s_T\}$ maximizing

$$\begin{aligned} P(\mathbf{X}, S) &= \prod_{t=0}^{T} P(s_t = j, \mathbf{x}_t | s_{t-1} = i) \\ &= \prod_{t=0}^{T} P(\mathbf{x}_t | s_t = j) P(s_t = j | s_{t-1} = i) \\ &= \prod_{t=0}^{T} \frac{P(s_t = j|\mathbf{x}_t) P(\mathbf{x}_t)}{P(s_t = j)} P(s_t = j | s_{t-1} = i) \\ &\sim \prod_{t=0}^{T} P(s_t = j|\mathbf{x}_t) \frac{P(s_t = j | s_{t-1} = i)}{P(s_t = j)} \end{aligned} \quad (1)$$

The probabilities $P(s_t = j|s_{t-1} = i)$ and $P(s_t = j)$ are easy to determine but for the determination of $P(s_t = j|\mathbf{x}_t)$ we have investigated two approaches. In the first one, a phone network (an MLP) is trained to estimate the $P(s_t = j|\mathbf{y}_t)$ with $\mathbf{y}_t$ being the output of the PHF detector. The desired $P(s_t = j|\mathbf{x}_t)$ are then substituted by the phone network outputs. In the second approach, $P(s_t = j|\mathbf{x}_t)$ is derived directly from $\mathbf{y}_t$. However, we argue that negative features form a majority and moreover, they are likely to be highly correlated (their y_{ti} are all small). Therefore, it is better not to take them into account, so as to prevent them from overruling the contributions of the positive features. Consequently, if the canonical features of state j are denoted as f_{ji} and if N_{pj} of them are equal to 1, we compute

$$P(s_t = j|\mathbf{x}_t) = \left[\prod_{f_{ji}=1} P(f_{ji}|\mathbf{x}_t) \right]^{\frac{1}{N_{pj}}} = \left[\prod_{f_{ji}=1} y_{ti} \right]^{\frac{1}{N_{pj}}} \quad (2)$$

The geometrical mean is obviously there to cope with the unequal number of positive features per state.

The admissible state sequences are obtained by expanding the phonemic transcription of the utterance down to the unit level, and by replacing each unit by a state. The linear automaton obtained in this way is then supplemented with skip arcs so as to model possible deviations from the priviliged state sequence. The phonemic transcription is either a manual or an automatic one that is retrieved from the orthography by means of a pronunciation lexicon. In the latter case we also allow parallel branches to accommodate different pronunciations of e.g. homonyms.

5.2. Experimental evaluation

The aligner was evaluated on the TIMIT core test set (24 speakers times 8 sentences) using a phone inventory of 48 phones (defined in [17]). We count as errors unit deletions, insertions and substitutions, and far errors referring to boundaries which are more than 20 ms off the manual boundary positions. The error rates for two probability computation strategies and two types of linguistic input are listed in Table 2. Apparently, the alignment is much more reliable when a manual phonemic transcription is available. The extra phone network does not outperform the simple model (there is only a small improvement when starting from a manual phonemic transcription).

For comparison we have also constructed a state-of-the-art HMM aligner with context-dependent phoneme models (triphones). Such a system performs a segmentation into phonemes and the evaluation has to be performed at the phoneme level too (42 phonemes). Table 3 shows that our system provides state-of-the-art segmentation and labeling performances, and thus, a good starting position for the construction of pronunciation proficiency scores.

prob comp	linguistic input	err (%)	del (%)	ins (%)	far (%)	sub (%)
simple model	ort	39.6	10.3	8.3	5.8	15.1
	phon	24.2	7.5	6.8	6.3	3.5
phone network	ort	40.0	10.8	7.7	7.3	14.2
	phon	22.2	7.1	4.9	6.9	3.3

Table 2. *Evaluation of segmentation and labeling for two systems and two types of linguistic input (core test set, 48 phonetic units)*

prob comp	linguistic input	err (%)	del (%)	ins (%)	far (%)	sub (%)
simple model	ort	42.1	11.9	9.2	7.3	13.7
	phon	28.1	8.6	7.9	8.0	3.6
HMM system	ort	48.2	8.8	12.3	12.7	14.4
	phon	32.9	8.0	8.8	12.8	3.3

Table 3. *Comparison of our aligners with an HMM-based system (core test set, 42 phonemes).*

6. PRONUNCIATION PROFICIENCY SCORING

Based on the computed alignments of his/her utterances we now characterize the pronunciation proficiency or the goodness of pronunciation (GOP) of a speaker. Therefore we first consider a GOP per phoneme, defined as the mean posterior probability of the correct label in the frames assigned to that phoneme. Then we analyze the posterior probabilities of the different PHFs in the frames assigned to a 'bad' phoneme in the hope to get information about the nature of the problem.

In a first experiment we assessed 20 speakers from the WSJ corpus: 10 native speakers (H2 test set) and 10 non-native speakers (S3 test set). The linguistic input was orthography and the CMU pronunciation dictionary. The non-native speakers were divided into 3 groups according to their native langauge (L1) (see Table 4).

group	speaker	L1	native country
S	4nd	Spanish	Argentinia
	4nh	Spanish	Israel
	4nm	Spanish	Nicaragua
F	4ne	French	France
	4nf	French	France
D	4ni	Danish	Denmark
	4nl	German	Germany

Table 4. *Non-native speakers and their mother tongue (L1)*

Per non-native speaker and per phoneme we counted the number of times a native speaker yielded a higher GOP for this phoneme, and we recorded for each speaker his/her bad phonemes as those for which this number was equal to 10 (the maximum). From Table 5 it appears that for every speaker a number of phonemes with a pronunciation deficiency can be identified. Moreover, the results confirm the expectation that, on average, Spanish and French speakers have a heavier accent than Danish/German speakers.

In a second phase we computed, per bad phoneme and per positive feature of that phoneme, the mean posterior of the feature, and

group	speaker	L1	bad phonemes
S	4nd	Spanish	ae, ih, d, g, l, n, iy
	4nh	Spanish	ch, eh, ih, ay, f, k, l, m, n, iy, dh, v, ow hh
	4nm	Spanish	ih, zh, d, f, l, iy, r, dh, s, v, z, hh
F	4ne	French	jh, aa, ih, d, l, m, n, iy, r, z
	4nf	French	aa, ae, ao, ih, zh, d, l, iy
D	4ni	Danish	jh, ih, er, t, ow, z,
	4nl	German	ih, f, l, iy, r, s

Table 5. *Detected bad phonemes per non-native speaker*

spkr	voic	vow	mid-high	front	voic	lat	alv
4nd	2	9	10	10	10	9	10
4nh	5	9	10	10	10	10	10
4nm	7	9	10	0	10	10	10
4ne	7	9	10	10	10	10	10
4nf	5	9	10	5	10	10	10
4ni	5	9	10	8	9	9	10
4nl	9	9	10	4	10	9	10
	phoneme /ih/				phoneme /l/		

Table 6. *Problematic features of phonemes /ih/ and /l/*

we counted the number of times a native yielded a higher value. For the phonemes /ih/ and /l/ appearing in the bad phoneme lists of most speakers these numbers are listed in Table 6. Aapparently, the system concludes that most of the non-native /ih/ sounds are produced at a higher position in the vowel triangle. This is confirmed by informal listening. The non-native speakers seem to produce a more /iy/-like sound. For the phoneme /l/, all the features seem to contribute similarly to the pronunciation deficiency.

7. CONCLUSIONS

We proposed a novel system for assessing the pronunciation proficiency of atypical speakers (non-natives, deaf, dysarthric, etc.). The basic assumption is that mapping the acoustics to a phonological feature space is a sensible step towards the retrieval of meaningful feedback on the gravity and nature of pronunciation deficiencies. Therefore, our system maps the traditional acoustic features (MFCCs) to phonological features, performs a segmentation and labeling of the utterance on the basis of these features and computes phonemic as well as phonological goodness scores to characterize the pronunciation proficiency of a speaker.

The results provided thus far demonstrate that state-of-the-art segmentation and labeling performance is obtained, that phonemes with deficient pronunciations can be detected in non-native speech and that it also seems possible to attribute these deficiencies to particular phonological features. Obviously, more work is needed to establish how well the discovered deficiencies correlate with human ratings of the assessed speakers.

8. ACKNOWLEDGMENTS

This work was supported by Flemish Institute for the Promotion of Scientific and Technical Research in the Industry (contract SBO/40102).

9. REFERENCES

[1] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. ICASSP*, 1997, pp. 1471–1474.

[2] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. ICSLP*, Philadelphia, USA, 1996, pp. 1457–1460.

[3] C. Cuchiarini, F. De Wet, H. Strik, and L. Boves, "Assessment of dutch pronunciation by means of automatic speech recognition technology," in *Proc. ICSLP*, 1998, pp. 1739–1742.

[4] K. Truong, A. Neri, C. Cuchiarini, and H. Strik, "Automatic pronunciation error detection : an acoustic-phonetic approach," in *Proc. of the InSTIL/ICALL Symposium*, 2004, pp. 135–138.

[5] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Automatic pronuciation error detection and guidance for foreign language learning," *Proc. ICSLP*, pp. 2639–2642, 1998.

[6] A.A. Wrench and W.J. Hardcastle, "A multichannel articulatory speech database and its applications for automatic speech recognition," in *5th Seminar on Speech Production*, Kloster Seeon, Bavaria, 2000, pp. 305–308.

[7] B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Turkey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," in *JASA*, 1978, number 63 in 5, pp. 1535–1555.

[8] J. Schroeter and M.M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," in *IEEE Trans. on SAP*, 1994, number 1 in 2, pp. 133–149.

[9] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," in *Computer Speech and Language*, 2000, number 14 in 4, pp. 333–353.

[10] R. Jacobson, G.M.C. Fant, and M. Halle, "Preliminaries to speech analysis: The distinctive features and their correlates," *MIT Press*, 1952.

[11] N. Chomsky and M. Halle, "The sound pattern of english," in *MIT Press*, 1968.

[12] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic bayesian networks," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 2529–2532.

[13] S. Chang, S. Greenberg, and M. Wester, "An elitist approach to articulatory-acoustic feature classification," in *Proc. Eurospeech*, Aalborg, 2001, pp. 1725–1728.

[14] M. Wester, S. Greenberg, and S. Chang, "A dutch treatment of an elitist approach to articulatory-acoustic feature classification," in *Proc. Eurospeech*, Aalborg, 2001, pp. 1729–1732.

[15] J.P. Martens and N. Weymaere, "An equalized error backpropagation algorithm for the on-line training of multilayer perceptrons," *IEEE Trans. on Neural Networks*, vol. 13, pp. 532–541, 2002.

[16] K. Kirchhoff, "Robust speech recognition using articulatory information," in *PhD Thesis*, Universität Bielefeld, 1999.

[17] K.F. Lee and H-W. Hon, "Speaker-independent phone recognition using hidden markov models," in *IEEE Trans. on ASSP*, 1989, number 11 in 37, pp. 1641–1648.