Automatic Detection of Mispronunciation in non-native Swedish Speech

Philippe Langlais, Anne-Marie Öster, Björn Granström
CTT (Center for Speech Technology) – TMH/KTH
SE-10044 Stockholm, Sweden
www.speech.kth.se

Abstract

This contribution presents part of the work initiated at CTT on the development of speech technology to assist non-native speakers in learning Swedish. This study mainly focuses on the automatic evaluation of mispronunciations at a phonetic level. We describe a new database we have collected for this work. Then we report the reliability of several phonetic scores to locate automatically segmental problems in student utterances.

1 Introduction

One of the greatest problems in the integration of immigrants is the language barrier. Even when having a fairly good knowledge of a foreign language, most adults will not lose their foreign accent, which represents a hurdle to social relation and employment. In the last decade speech technology improvements has opened new possibilities in interactive language teaching systems [9, 5, 8]. Very recently a growing number of studies addressed the problem of rating automatically non-native speakers providing measurements that correlate with human judgments [2, 6, 4, 10]. These last studies shows that rating a speaker on a 1 to 5 scale can be achieved (providing enough sentences) with a performance close to the one of human experts. The results reported on the rating of each sentence separately are however not convincing. This work address the automatic rating of individual phonemes, which is still a challenging task.

2 Database

2.1 Speech Material

A database composed of 21 speakers (6 females) has been collected for research purposes. The speakers (all of them have an academic degree in engineering from their respective university) come from different countries (mainly Cuba, Peru, Bosnia, but also Germany, Saudi Arabia, Russia, etc.). Each one read both a text and isolated words. The reading material was selected by teachers with regard to several constraints: 1) all Swedish vowels (long and short) are present at least once in the database; 2) all consonants are spoken in all phonotactic positions (initial, medial and final).

The text was selected to provide easy reading for low-skilled students in Swedish (17 sentences with an average of 17 syllables each). 110 isolated words were also produced by each speaker (59 mono-, 50 bi-, and 1 trisyllabic words); 18 of them were uttered twice. The students were first asked to read silently both the words and the text, in order to allow them to ask questions about the possible pronunciation problems they might encounter. Then, the recording session started during a real examination in a classroom and was made on an analogue cassette recorder.

2.2 Transcription

All this material is now being transcribed by experts. A user interface has been developed for this purpose, thus providing experts with classical and specific annotation tools (automatic phonetic alignment, pitch extractor, display of the signal, common errors, etc.). As pointed out by [1], very little is known about the human scores considered as a reference in previous studies. Therefore, we asked the experts to be aware of the different parameters that can influence their rating. For instance, we ask each expert to distinguish between prosodic and phonetic quality during the entire transcribing process which consist of the following tasks:

- Rating the prosodic, the phonetic and the global quality of each item (word or sentence) using a 1 to 5 scale (“unintelligible” to “sounds like Swedish”).
- Identifying and rating each prosodic deviation. Phonetic deviations were identified by modifying phonetic symbols in a standard phonetic transcription which was provided for each item using a text-to-phoneme processor [3]. Each reference has been manually checked and corrected before the transcription process. Each phonetic deviation was rated using a 1 to 5 scale (“horrible” to “not really deviant”).
- Rating globally each session on a 1 to 5 scale. A session is the pronunciation of either the list of words or the text. This score is intending to reflect the global skill of the student. We make the assumption that a given student may be differently skilled reading sentences and words.

2.3 Expert feedback

Three experts are involved in the transcribing process. Two of them have a strong background in phonetic science. All are involved in the field of foreign language learning.

First, it appears that the transcribing task is much more complex and effort-consuming than what was initially expected. For instance, experts often feel the need to come back several times to previous items or sessions to insure a consistency in ratings.

For the time being, only one expert has transcribed enough material to be considered in this study. Therefore we limit this investigation to 5 text sessions (about 800 phonemes each) and 6 word-sessions (close to 400 phonemes each).

This expert pointed out that rating sentences is easier than rating isolated words and expressed a difficulty to rate globally each speaker session. Looking at the correlation between the three different scores that the expert is asked to give for each item, we observe for the text material that the global rating seems to be more correlated to the prosodic rating ($r_{xy} = 0.82$) than to the phonetic one ($r_{xy} = 0.70$). It will be part of our future work to check the consistency among the different experts' judgments. However, we can already remark that comparing phoneme-score algorithms with human judgment at the speaker level may be a questionable task.

3 Experiments

In this work we investigate the usefulness of the acoustic scores provided by a speech recognizer to rate the phonetic quality of non native speaker at the phoneme level. Therefore only the phonetic scores provided by experts are taken into account in the following. For that purpose we used the set of context independent phoneme models that is used in the Waxholm recognition system [7], despite the fact that better recognition rates are obtained with contextual models. Moreover Young & Witt [10] reported higher accuracy on a task of phoneme quality acceptance/rejection, using monophones models.

3.1 Global recognition rates

The basic assumption behind the phonetic scoring algorithms proposed in previous studies is that the acoustic scores computed by a speech recognizer trained on native speech should correlate to the quality of non-native speech. As a first crude test of this assumption, we ran two small recognition experiments: the former is the recognition of isolated words (90 different words, only one possible pronunciation provided for each), in the latter, we used a grammar set as a loop of all possible words in the text material (141 words plus silence considered as a normal one) which was applied to the recognition of the sentences produced. As our intention is to investigate the acoustic scores only, we didn't give the recognizer any grammar constraints (e.g. word bigrams). The accuracy rates at the phonetic level as well as the experts' ratings are reported for each speaker in figure 1. In both recognition tasks, the accuracy is highest for the native speaker (Nat). The correlation between the accuracy and the average phonetic rates is 0.58, while the correlation between the accuracy and the global rates is 0.49.

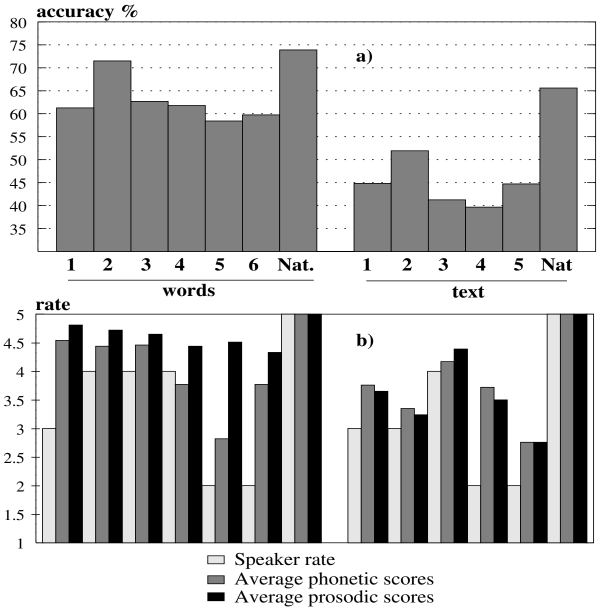

Figure 1: *Accuracy rates at the phonetic level (top) of each small-scale recognition task: the isolated* word *material for the isolated words material (from left to right, speakers 1 to 6 plus a native one) and the* text *material (speakers 1 to 5, plus a native one). Plot b) reports the global speaker score, and the two average experts' scores (over the item of a session): the average prosodic score and the phonetic score.*

3.2 Measurements

We investigate two kinds of phonetic scores both computed in an HMM paradigm: log-likelihood scores and log-posterior probability scores. We decided not to look at information such as segment duration scores, as we consider it to belong to the suprasegmental level, which we plan to approach later. Furthermore Kim et al. [4] reported the bad performance of the duration score to rate quality of specific phones.

Log-Likelihood scores For each phone segment i (of d frames: $[t_0, t_0 + d[$), the log-likelihood score is defined as l where $p(y_t|q_i)$ is the probability that the observation vector at time t (y_t) is generated via the phonetic model q_i:

$$l = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log p(y_t|q_i) \quad (1)$$

The log-likelihood score can be directly taken as it is computed during the viterbi forward pass. It means in this case that we don't have the control of which state (if more than one) of the model under consideration has been used for the computation at time t. Therefore we computed $p(y_t|q_i)$ as:

$$p(y_t|q_i) = \max_{j=1}^{n_i} b^x_{i_j}(y_t) \qquad (2)$$

where $b^x_{i_j}$ stands for the density probability function (typically a mixture of gaussians) associated to the jth state of the model q_i (which totalizes n_i states). The symbol x represents the number of coefficients used to compute b_{i_j} for the observation y_t (39 coefficients: 13 first order mfcc coefficients plus Δ and $\Delta\Delta$). Practically, we tested three values of x: 13, 26 and 39 yielding three reported log-likelihood measurements (namely l_{13}, l_{26} and l_{39}).

Log-posterior probability scores Regarding previous studies, it seems that HMM-based likelihood scores are poor predictors of the phonetic quality (at least at the sentence level). This can be explained by the fact that non-discriminative criteria are still widely used during the training process of an HMM paradigm. A solution here is to compute the log-posterior probability score of a phone segment i as the average (over time) of the frame based posterior probability of the phone q_i at time t. N stands for the number of phonetic models (namely 47 here) and $P(q_i)$ is the prior probability of the phone class q_i:

$$\rho = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log \frac{p(y_t|q_i)P(q_i)}{\sum_{j=1}^{N} p(y_t|q_j)P(q_j)} \qquad (3)$$

Here again, we can compute this score in several ways. For instance [10] suggested to compute it assuming all phones are equally likely (that is H_1: $P(q_i) = P(q_j), \forall i, j$) and that the sum in the denominator can be approximated by its maximum (H_2). We report results of two scores ρ_1 (H_1 and H_2) and ρ_2 (H_2).

3.3 Methodology

To evaluate the efficiency of each score we first aligned the expected phonetic string (time aligned) with the expert transcription. As pointed out by [10], a student can make different mistakes with which a system has to cope: insertions, deletions and substitutions. We defined a dynamic programming scheme suitable for the present task (allowing for instance 2-1 mappings which are common with non-native utterances). We however limit this study to the most common and easier errors to handle: the substitutions (1-1 mappings). Among these errors, we focussed on the vowels substitutions.

The trade in previous studies was to measure the reliability of a scoring algorithm, computing the correlation between its scores and the experts' rates. We think that this measurement should not be given alone. Collecting feedback from experts, it appears first that a given phonetic deviation can be rated differently depending on where it occurs (common words, function/grammatical words, etc.). This should slightly affect the correlation results. More important is the fact that this measurement depends of the number of observations (in our case phonemes) that have been rated not deviant. This can be seen in figure 2. This suggests that correlation coefficient should not be given alone. Therefore, each score is evaluated on its effectiveness to separate phonemes in two classes: the deviant ones and the others. The efficiency of each score is given by the classical precision (P) and recall (R) rates that are summarized for convenience in the following by the F-measure f: $f = 2 \times \frac{P*R}{(P+R)}$. In other words, we argue that it is more informative to evaluate separately the two parts involved in such studies: the mispronunciation detection part and the rating one.

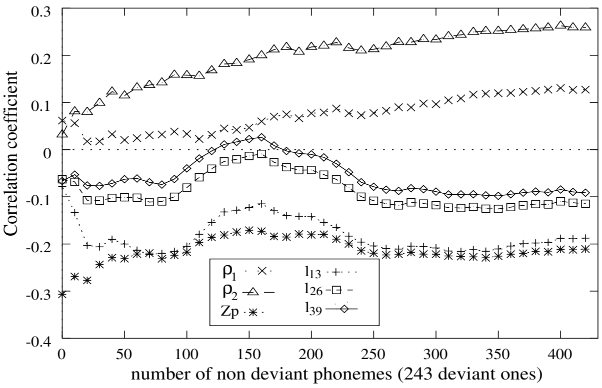

Figure 2: *Evolution of the correlation coefficient of different scoring algorithms as a function of number of non deviant observations. Z_p is a normalized log-likelihood score (see text explanations).*

3.4 Results of Individual Ratings

The results of the scoring algorithms described in section 3.2 are reported in table 3.4 and invite some comments. First of all, we must observe that the correlations between these scores and the experts' rates are rather small (see also figure 2) and that a binary decision of the quality of a vowel (good/deviant) could not be achieved with a rate much more higher than 0.5 (both for the sentences and for the words). Several explanations can be advanced to explain this phenomena. First the HMM set used here is probably not accurate enough (2 gaussians per state, due to real time computation constraints). However, we checked manually a subset of the alignments of the expected phonetic string to the signal without noticing important errors. It is part of our future work to investigate more accurate sets of models. A more important reason can be found in the inventory of the errors considered in this experiment. Most of them are fine deviations, mainly mismatch between the phonetic quality of an accented vowel and its non accented counterpart (e.g. [o]/[ɔ] substitution).

In order to verify the behavior of the scoring algorithms when dealing with more severe substitutions, we ran a second experiment on an artificial database. Instead of

Table 1: *F-measures (percentage values) and correlation coefficients between different scoring algorithms and the expert phonetic rates observed for two experiments: isolated words (690 vowels, 243 deviants ones) and sentences (917 vowels, 308 deviants)*

	words		sentences	
Score	cor.	$f(\%)$	cor.	$f(\%)$
l_{39}	-0.08	52.2	-0.04	50.4
l_{26}	-0.10	52.3	-0.08	50.5
l_{13}	-0.17	54.5	-0.20	55.0
ρ_1	0.11	52.1	0.16	51.3
ρ_2	0.25	52.1	0.22	53.6

recording a specific erroneous database such as in [10], we modified the phonetic reference in around 30% of the non-deviant vowels. The modifications were crude ones such as [*i*]/[*a*] substitutions. The results of this experience are reported in table 3.4. We can observe that in such conditions, the scoring algorithms behaved as expected and that vowels can be well identified as deviant or not. The results are higher on isolated words than on sentences which gives an indication of the effect of the time alignment process on the results.
In both experiments, we observe that computing the log-likelihood using only the first order coefficients gives better results than taking into account the derivative ones. Looking only at the correlation, the log-posterior probability scores seems more efficient, which has also been observed in other studies. In the first experience, we observe that ρ_1 (log-posterior probability score with all models equally likely) is slightly less efficient than ρ_2 (using a unigram model), which is not surprising since the acoustic information alone seems not strong enough here to make robust decision.

Table 2: *F-measures (percentage values) computed for different scores on two artificial databases:* words *(448 vowels, 151 deviant ones) and* sentences *(614 vowels, 204 deviant ones).*

	l_{39}	l_{26}	l_{13}	ρ_1	ρ_2
words	85.5	89.1	91.2	94.6	94.4
sentences	67.9	75.4	82.2	83.2	82.2

4 Conclusion

We presented a new database Swedish pronounced by non-native students, suited for studies of speech technology in language learning. We discussed some points of an evaluation protocol and we applied different scores to the tasks of locating and rating specific deviant vowels. Further experiments should be carried out with more accurate HMM-models to reach a better level of performance.

Acknowledgments

This work was partly founded by CTT, Centre for Speech Technology. We are indebted to the Unit for Languages and Educational Research and Development (KTH) that kindly gave us material about their L2 students pronunciation training.

References

[1] C. Cucchiarini and L. Boves. Automatic assessment of foreign speakers' pronunciation of Dutch. In *5th European Conference on Speech Communication and Technology*, volume 2, Rhodos, 1997.

[2] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. Automatic pronunciation scoring for language instruction. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1471–1474, Munich, 1997.

[3] J. Gustafson. A Swedish Name Pronunciation System. Licenciate thesis, TMH, KTH, 1996.

[4] Y. Kim, H. Franco, and L. Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In *5th European Conference on Speech Communication and Technology*, volume 2, pages 645–648, Rhodos, 1997.

[5] J-P. Lefèvre and M-G. Di Benedetto. Macro and micro features for automated pronunciation improvement in the SPELL system. *Speech communication*, 11(1):31–44, 1992.

[6] O. Ronen, L. Neumeyer, and H. Franco. Automatic detection of mispronunciation for language instruction. In *5th European Conference on Speech Communication and Technology*, volume 2, pages 649–652, Rhodos, 1997.

[7] N. Ström. Continous speech recognition in the Waxholm dialogue system. TMH-QPSR 4/1996.

[8] Y. Taniguchi, A.A. Reyes, H. Susuki, and S. Nakagawa. An English conversation and pronunciation CAI system using speech recognition technology. In *5th European Conference on Speech Communication and Technology*, volume 2, Rhodos, 1997.

[9] C.S. Watson, D.J. Reed, D. Kewley-Port, and D. Maki. The Indiana speech training aid (ISTRA): Comparisons between human and computer-based evaluation of speech quality. *Journal of Speech and Hearing Research*, 32:245–251, June 1989.

[10] S. Witt and S. Young. Language learning based on non-native speech recognition. In *5th European Conference on Speech Communication and Technology*, volume 2, pages 633–636, Rhodos, 1997.