PHONETIC-LEVEL MISPRONUNCIATION DETECTION 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 the CTT for the development of speech technology to assist non-native speakers learn Swedish. This study focuses mainly on the automatic location of mispronunciations at a phonetic level. We first describe the database we created for this work and then report on the reliability of several phonetic scores to automatically locate segmental problems in student utterances.

1. INTRODUCTION

Over the last decade, advancements in speech technology have opened up new possibilities for interactive language teaching systems [1,7,10,11]. And more recently, a growing number of studies have addressed the problem of automatically rating non-native speakers by providing measurements that can be correlated with human judgment [3,5]. These studies show that rating a speaker on a 5-point scale can be achieved with performance inversely proportional to the size of the speech-unit rated. In other words, speech technology is mature enough to grade a speaker globally and seems capable of grading individual sentences. Despite an interesting implementation [8], it is not yet clear whether speech technology is suited to rate individual words. Furthermore, few studies have reported results on grading individual phonemes [5,6,12,13]. Although this task is quite challenging, we do believe that a system capable of grading pronunciation at a phoneme level is a prerequisite for a useful stand-alone pronunciation trainer. Therefore we address in the following the automatic rating of individual phonemes using speech recognizer outputs.

2. DATABASE

2.1. Speech Material

Twenty-one non-native speakers of Swedish, including 6 females, participated in this research. All hold an engineering degree from their respective universities. Each speaker was asked to read isolated words as well as a full text. The reading material, as chosen by teachers of Swedish, included 1) all Swedish vowels (long and short) at least once; 2) all consonants in all initial, medial and final position. The text (17 sentences with an average of 17 syllables each) was selected to provide easy reading for low-skilled students of Swedish. Each speaker pronounced 110 isolated words (59 mono-, 50 bi-, and 1 tri-syllabic words), with 20 of them being uttered twice. First, the students were asked to read the words and the text silently and then given the opportunity to ask questions about the possible pronunciation problems they might encounter. Their utterances were recorded on analog tape during a classroom test.

2.2. Transcription

Experts then transcribed the material. A user interface was developed for this purpose, thus providing experts with classical and specific annotation tools (automatic phonetic alignment, pitch extractor, signal display, common errors, etc.). As pointed out by [2], very little is known about the human scores used as references in previous studies. Therefore, we asked the experts to be aware of the different parameters that could influence their ratings. For instance, each expert was asked to distinguish between prosodic and phonetic quality during the entire transcription process. Furthermore, the expert modified the standard phonetic transcription provided for each item by a text-to-phoneme processor [4] to reflect the phonetic deviation. Each phonetic deviation was rated on a scale of 1 to 5 (from “horrible” to “not really deviant”). More information on the transcription process is given in [6]. For the time being, only one expert has transcribed enough material to be considered in this study. This expert has a strong background in phonetic science and is involved in foreign language learning. She has transcribed 8 text-sessions (about 800 phonemes each) and 6 word-sessions (close to 400 phonemes each).

2.2. Expert feedback

It appears that the task of transcribing is much more complex and time-consuming than was initially expected. To ensure consistent ratings throughout her work, the expert felt that she often needed to go over several times item which had already been dealt with and even review work she done in previous sessions. The expert also pointed out that rating sentences is easier than rating isolated words and expressed difficulty in giving an overall rating to each speaker session. As part of our future research, we will check the consistency within and between the different experts' judgments.

2.4. Error analysis

The greater the segmental error, the easier it is to detect [6]. Hence, we will discuss the distribution of vocalic errors observed in our corpus regardless of the mother-tongue of the speaker. We must first point out that a specific error doesn't necessarily receive the same rating. For instance, the vowel [ɑː] pronounced as [a], was graded three times as 1, twice as 2 and once as 3. This is due to the fact that the expert's judgement was influenced by the context in which the error occurred (syllable, function/grammatical words, etc.). The influence of the context, and what is actually understood by context, is still open for debate and cannot be investigated yet due to lack of data. However, it is a rather important issue given that the task consists in matching automatic scoring with the evaluations of experts. The distribution of vocalic errors is reported in figure 1. Each error is indexed as a single value (called sp-index) representing its position in the classical 3-dimensional feature-vowel production space: the degree of opening (close, close-mid, open-mid and open), the front/back dimension (front, central and back) and the lip-rounding dimension (rounded or not). Each deviation of one unit in any of these three dimensions scores a 1 and the sp-index of an error is computed as the sum of the deviations for each dimension. For instance, the error [a]/[ɔ] is indexed as 3 since these two vowels differ by two units in the front/back dimension (front vs. back) and by one unit in the opening dimension (open vs. open-mid). Looking at the error distribution, we observe that most of the errors in our corpus deviate in only one of the three feature-dimensions, as is the case for the close-mid/open-mid error [o]/[ɔ].

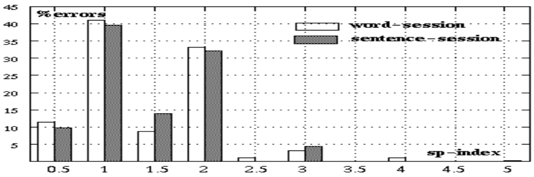

Figure 1: Distribution of vocalic errors as a function of sp-index (see text explanations).

3. EXPERIMENTS

In this work, we investigate the usefulness of acoustic scores provided by a speech recognizer to identify whether vowels are deviant or not. For that purpose, we used several sets of context independent phoneme models trained on a native clear-speech corpus of 2100 sentences. Two different types of feature vectors were evaluated: standard static Mel frequency cepstrum coefficients (MFCC) and dynamic mean cepstrum subtraction Mel frequency cepstrum coefficients. The latter should perform better when there is background noise or when there is a significant mismatch between the training and target data, which is the case with our non-native corpus (clear speech vs. cassette recorder). Due to practical issues, we have not yet investigated contextual models. However, when they used multiple mixture component monophones models, Young & Witt [12] reported a higher degree of accuracy for phoneme quality acceptance/rejection.

3.1. Global recognition rates

The basic assumption behind the phonetic scoring algorithms proposed in previous studies is that one should be able to evaluate the quality of non-native speech by acoustic scores computed by a speech recognizer trained on native speech.

As a first crude test of this assumption, we ran a small-scale isolated words recognition experiment (a loop of the 90 different words of the word-sessions described in section 2.1, with only one possible pronunciation provided for each word). The recognition rates for the static sets of models as well as the experts' ratings are reported for each speaker in figure 2. Contrary to our expectations, the recognition rates observed with dynamics sets of models are always lower (a loss of accuracy exceeding 15%). We observe that the word recognition rate is poorest for the two lowest ranking speakers, and that the first-ranked speaker is the one that is best recognized.

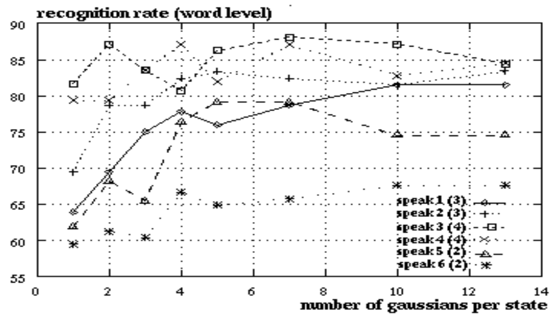

Figure 2: Word recognition rates measured for six speakers as a function of number of gaussians per state. The number in parentheses reports the speaker rate given by an expert.

3.2. Measurements

Further to previous work, we compared two types of phonetic scores, log-likelihood scores and log-posterior probability scores, which are both computed in an HMM paradigm.

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:

$$l = \frac{1}{d} \cdot \sum_{t=t_0}^{t_0+d-1} \log(p(y_t / q_i)) \quad and \quad p(y_t / q_i) = \mathop{\nabla}_{j=1}^{n_i} b_{i_j}^x y_t$$

where $b_{i_j}^x$ 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). ∇ stands for an operator (maximum or average). The symbol x represents the number of coefficients used to compute $b_{i_j}^x$ for the observation y_t. Practically, we tested two values of x: 13 (only first mfcc coefficents) and 39 (first coefficients plus Δ and ΔΔ ones).

Computing a frame by frame log-probability as described here provides output scores for each phoneme, and this correlates more or less to the log-probability score directly computed during the forward viterbi path. The correlation coefficients range from 0.5 to 0.9 depending on the speakers tested and the set of models used. One possible explanation is that in the frame-by-frame (FBF) computation mode, the order in which different states are considered is not necessarily the order in which they occur in time. Therefore we also endeavoured to compute log-likelihood scores by running a viterbi alignment on each segment. The likelihood probabilty at time t is obtained inverting the viterbi path. In this paper, we will refer to this mode of computation as VIT as opposed to the FBF algorithm. We also investigated whether or not it was best to include the transition probability in the score calculated by the viterbi path, and found that, for this task at least, it didn't make a significant difference.

Log-posterior probability scores

Based on previous studies, it seems that HMM-based log-likelihood scores are poor predictors of the phonetic quality (at least at the sentence level). This can be explained by the fact that non-discriminating criteria are still widely used during the training process of an HMM paradigm. The log-posterior probability score of a phone segment *i* is computed 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 53 here) and $P(q_i)$ is the prior probability of

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

the phone class q_i :

We computed this score assuming all phones equally likely.

3.3. Evaluation

Due to time constraints, the expert transcription is not time aligned. Thus, in order to evaluate the efficiency of each score, we aligned the expected phonetic string (which is time aligned) with the expert transcription. 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 in non-native utterances). We limited this study to the most common and easiest errors to handle, namely substitutions (1-1 mappings). Among these errors, we focussed on vowel substitutions despite believing, given our corpus, that spotting consonant errors could be an easier task. As a matter of fact, the spelling of some Swedish words often causes beginners to make significant consonantic errors. For instance the letter **k** is pronounced [k] when it is followed in the same syllable by **e, i, y, ä** or **ö**, but is pronounced [ç] either. In previous studies the reliability of a scoring algorithm was measured by computing the correlation between its scores and the experts' ratings. As discussed in section 2.4, it appears that a given phonetic deviation can be rated differently depending on the context in which it occurs. This should affect the correlation results slightly. An artefact which can influence the interpretation of the results was also discussed in [6]. In this study therefore we evaluate each score on its ability to locate deviant phonemes. The efficiency of each score is measured according to the precision (*P = bad / (bad+GOOD-good)*) and recall (*R = bad / BAD*) rates. For the sake of convenience, we will use the F-rate (*F = 2× (P·R)/(P+R)*) where *BAD* and *GOOD* (resp. *bad* and *good*) stand for the number of phonemes that have been respectively rejected and accepted by an expert (resp. by the system). Note that the score accuracy (*SA*: rate of phonemes well identified) and the false acceptance rate (*FA*: rate of deviant phonemes that have been accepted) as used in [12] is less informative for corpora containing fewer deviant phonemes (then *SA* is basically higher than *P*).

3.4. Results of Individual Ratings

We tested 372 scoring machines which combined different factors such as the number of coefficients of the input speech vector (13 or 39), the kind of set of models (static or dynamic, from 1 to 13 gaussians), the score used (log-likelihood or posterior scores), etc. Using a native corpus, each scoring machine was automatically assigned a set of threshold values for each vowel encountered. The distribution of F-rates observed for those scoring machines is reported in figure 3. Basically, two types can be distinguished: those with performances around 0.15 and those rating around 0.4. The best machines obtain a performance above 0.5 which means, roughly speaking, that fifty percent of the vowels identified as deviant were really deviant, and that around half of the deviant vowels were identified. This is obviously not good enough for a realistic application despite the fact that this approach outperforms a random scorer. Several reasons can account for that. First, as already discussed, most errors are due to slight deviations, mainly mismatches between the phonetic quality of an accented vowel and its non accented counterpart (e.g. [o:]/[ɔ] substitution). In a previous study [6], we ran an experiment on an artificial database. Modifications such as [i]/[a] substitutions were made to the phonetic reference of 30% of the non-deviant vowels. Under such conditions, the scoring algorithms behaved as expected and vowels could indeed be labelled as deviant or not. We obtained an F-rate of more than 0.94 on word-sessions with posterior scores computed with a set of 2-gaussian static models. Another reason that can explain the limit of these scoring machines is the difference in quality between the native speech database and the non-native one. As part of our future work, we plan to investigate the impact of noise reduction techniques on the results.

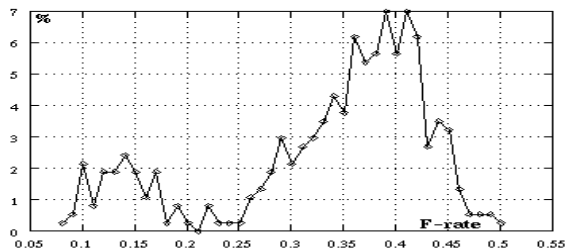

Figure 3: Distribution of the number of scoring machines investigated as a function of F-rate.

We will now discuss how the different factors influenced the performances observed. Figure 4 shows as a function of the F-rate, the ratio of systems that make use of a) the viterbi algorithm (VIT), b) the static set of models (STATIC), c) the 39 coefficients (COEF39) and d) log-likelihood scores (PROB). We observe (figure 4a) that the worst systems (20% of the total ones) represent around half of the systems which used log-likelihood score and a third of the systems using static monophones. This corroborates what has been observed in other studies: posterior scores are better predictors of phoneme quality than log-likelihood scores. With less influence, static models are worse predictors than dynamic ones mainly because of the mismatch in quality between the training and the testing data. Looking at a few of the better systems (figure 4b), we observe that they use neither log-likelihood scores nor static models. They all use the viterbi computation on input vectors of 39 coefficients. Despite the fact that the differences in performance are small, the scoring machine that outperformed the others uses 3-gaussian dynamic models, computing posterior scores on 39 coefficients. The performances observed on the sentence-sessions (1225 vowels, 269 deviant ones) range from 0.17 to 0.47. The best results for the word-sessions were obtained by systems using posterior scores in their calculations, and the worst results by those using log-likelihood scores. The best and worst performances for both sessions are reported in table 1

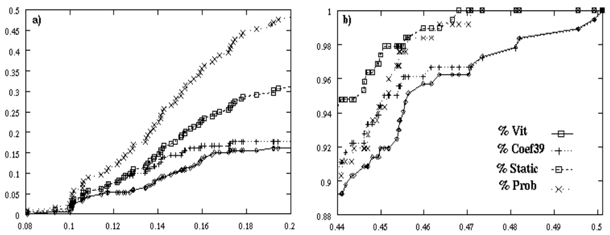

Figure 4: Distribution of several ratios characterizing the scoring machines investigates as a function of F-rate.

	Word-session (806/238)				Sentence-session (1125/269)			
	SA	FA	P	R	SA	FA	P	R
B	0.67	0.28	0.54	0.60	0.47	0.35	0.36	0.86
W	0.54	0.41	0.20	0.10	0.49	0.35	0.19	0.34

Table 1: Score accuracy (SA), false acceptance (FA), precision (P) and recall (R) rates obtained by the best scoring machine (B) and the worse one (W) for both word-sessions (806 vowels, 238 deviant) and sentence-session (1125 vowels, 269 deviant).

4. CONCLUSION

In this paper, we described a new database of Swedish as spoken by non-native students. This data is particularly well-suited for studies in speech technology as applied to language learning. We investigated and compared different ways of rating the acoustic quality of a phoneme. We must note that the lack of target human rates at the segmental level makes the interpretation of the results somehow tricky.

Acknowledgments

This work was funded, in part, by the CTT, Centre for Speech Technology. We are indebted to the Unit for Languages and Educational Research and Development (KTH) who kindly gave us material about their L2 students' pronunciation training. We would also like to thank Lucie Langlois for her useful comments.

5. REFERENCES

1. Bernstein J. “New uses for speech technology in language education”, *Proceedings of ESCA workshop STILL´98, Marholmen, Sweden*, 1998.

2. Cucchiari C. and Boves L. “Automatic assessment of foreign speakers’ pronunciation of Dutch”, *Proceedings Eurospeech ’97. Rhodes, Greece*, 1997.

3. Franco H., Neumeyer L., Kim Y. and Ronen O. “Automatic pronunciation scoring for language instruction", *Proceedings ICASSP’97, Munich, Germany*, 1997.

4. Gustafson J. *A swedish Name Pronunciation System*, Licenciate thesis, TMH, KTH, 1996.

5. Kim Y., Franco H. and Neumeyer L., “Automatic pronunciation scoring of specific phone segments for language instruction”, *Proceedings Eurospeech ’97. Rhodes, Greece*, 1997.

6. Langlais P., Öster A-M. and Granström B. “Automatic detection of mispronunciation in non-native Swedish speech”, *Proceedings of ESCA workshop STILL´98, Marholmen, Sweden*, 1998.

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

8. Neumeyer L. & al. “WebGrader™: A multilingual pronunciation practice tool”, *Proceedings of ESCA workshop STILL´98, Marholmen, Sweden*, 1998.

9. Ronen O., Neumeyer L. and Franco H. “Automatic detection of mispronunciation for language instruction”, *Proceedings Eurospeech ’97. Rhodes, Greece*, 1997.

10 Taniguchi Y., Reyes A.A., Susuki H. and Nakagawa S. “An English conversation and pronunciation CAI system using speech recognition technology”, *Proceedings Eurospeech ’97. Rhodes, Greece*, 1997.

11 Watson C.S., Reed D.J. Kewley-Port D. and Maki D. “The Indiana speech training aid (ISTRA): Comparison between human and computer based evaluation of speech quality”, *Journal of Speech and Hearing Research, 32:245-251*, 1989

12 Witt S., Young S. “language learning based on non-native speech recognition”, *Proceedings Eurospeech ´97, Rhodes, Greece*, 1997.

13 Witt S., Young S. “Performance measures for phone-level pronunciation teaching in CALL, *Proceedings of STILL´98, Marholmen, Sweden*, 1998.