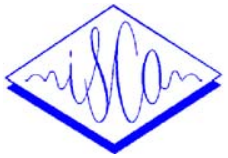

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AUTOMATIC DETECTION AND CORRECTION OF PRONUNCIATION ERRORS FOR FOREIGN LANGUAGE LEARNERS : THE DEMOSTHENES APPLICATION

*G. Deville**, O. Deroo*, H. Leich* S. Gielen** & J. Vanparys***

***Facultés Universitaires, 61 rue de Bruxelles – B-5000 Namur – Belgique*
**Faculté Polytechnique, 31 boulevard Dolez – B-7000 Mons – Belgique*

guy.deville@fundp.ac.be- deroo@tcts.fpms.ac.be - leich@tcts.fpms.ac.be
sofie.gielen@fundp.ac.be - johan.vanparys@fundp.ac.be

ABSTRACT

This paper accounts for the DEMOSTHENES application, an interactive tool for the correction of foreign language learners' pronunciation. After giving the didactic and technical arguments advocating for DEMOSTHENES, the authors describe the phase of creation and labelling of a sound database for the application language (Dutch). The methodolgy for the training of the recognizer is then discussed, as well as the pronunciation scoring paradigm. To conclude, the final application is described, together with preliminary but promising results.

1. INTRODUCTION

Acquiring a good command of spoken Dutch is a non-trivial task for most French speaking learners. Traditional audio-visual aids -- in the classroom or the language laboratory -- have shown their limitations in correcting pronunciation (lack of systematic feed-back in a non invidualized environment). On the other hand, current techniques in continuous speech processing make it possible to develop multimedia tools that analyse and correct the foreign language learner's pronunciation in a consistent and individualized approach. In this prospect, two Belgian research teams have joined their expertise in speech recognition (Polytechnique - Mons) and software development for foreign language learning (Namur University) to launch the DEMOSTHENES project.

DEMOSTHENES will result in a multimedia courseware for Dutch pronunciation, which detects and corrects the typical errors made by French speaking learners. Thanks to the well-known hybrid HMM/ANN systems which aim at combining Hidden Markov Models and Artificial Neural Network, the final product will identify pronunciation errors at the phoneme level.

Such a fine-grained approach distinguishes the DEMOSTHENES application from products currently available on the market, which provide feed-back in a graphic, non-linguistic format.

2. CREATION AND LABELLING OF A SOUND DATABASE FOR DUTCH

The DEMOSTHENES data base consists of isolated words, phrases and sentences that are representative of the pronunciation errors made by French-speaking learners [5] (e.g. language-specific phonemes without equivalent in French, assimilations, confusion between long/short vowels, etc.). About 25 different pronunciation difficulties are illustrated in a sample of several hundred items, pronounced by 135 (native and non-native) speakers of Dutch. Table 1 gives the number of speakers with gender, native vs. non-native origin. Table 2 indicates the age of the speakers.

	Female	Male
Native	34	33
Non-native	41	27

Table 1. Number of speakers with gender and native vs. non-native origin.

As DEMOSTHENES will address a wide range of learners (from beginners to more advanced learners), the phrases and sentences of the database have been carefully selected from the basic vocabulary of Dutch (2000 most frequent words). The contextual approach has been favoured to the detriment of a mere series of phonetic pairs to imitate. For practical reasons due to the size of the full database, three groups of speakers -- selected at random -- pronounced each a different overlapping

database subset of more managable size (about 350 items), as indicated in Table 3. Three speakers went through the full size database only for testing purpose.

Age	Number of speakers
-10	1
11-20	40
21-30	63
31-40	8
41-50	13
51-60	4
+60	6

Table 2. Age of the speakers.

.

Database	Number of items	Number of speakers
Subset a	349	44
Subset b	345	44
Subset c	346	44
Full database	593	3

Table 3. Number of items pronounced by speakers.

Basic phonetic units have been labelled in the specific context of DEMOSTHENES : an extended phonetic alphabet has been defined for the coding of the sound database, including erroneously pronounced phonemes. The treatment of Dutch pronunciation features is learner-dependent : are considered relevant only mistakes with the highest score of expectations to French speakers (due to contamination with the mother tongue). Other mistakes are not labelled as such, and are thus irrelevant in this context (due to their low expectation).

3. TRAINING OF THE RECOGNIZER

The recognition engine is being trained by the STRUT software (Speech Training and recognition Unified Tool [1]).

We train two different hybrid HMM/ANN systems, one for the native and one for the non-native speakers. Each system is being trained using an iterative process commonly used in speech recognition.

We use a first hand labelled segmentation in order to initialise our models and then use those models in order to re-label the native and non-native database. This process is iterated until the training process converge.

We decided to train two different MLPs (one for the native and one for the non-native speakers). The explanation of this choice is that some of the non-native (resp. native) speakers have probably produced, during the recording of the database, speech as a mixture of native and non-native phones. As the MLPs are locally discriminant, we would have had many problems in training an unique MLP to discriminate the same phones said by a native and a non-native speakers. The 3-layers MLP is trained by using stochastic gradient descent and relative entropy as the error criterion. A sigmoïde function is applied to the hidden layer and output units [5].

The network has an input layer of 234 units spanning a window of 9 frames, where each frame consists of cepstral parameters (log RASTA-PLP [6]), the Δcepstral parameters, the Δenergy and ΔΔenergy. The log-RASTA-PLP parameters have been chosen because of their relative robustness against noise and microphones change.

As we are working at the phoneme level, we define an output layer of 49 units, corresponding to one unit per phonetic class.

We extract 30 speakers from the native database and 30 speakers from the non-native database. We tried of course to have the same number of speakers in both databases and the same male-female repartition in each of them. This lead to approximately 25.000 sentences and 10 hours of speech.

The following table shows the recognition rate at the frame level on the training and cross-validation set for the native and the non-native models.

	Nbr Kframes	% Rec
Native Train	2898	80.2%
Non-native train	1379	76.7%
Native Cross	332	76.5%
Non-native Cross	141	73.8%

Table 4 : Phone Recognition Rate at the frame level with the MLPs trained on the native and non-native database and log-RASTA-PLP parameters.

As shown in Table 4, we are able to evaluate at each time, the probability of being in one particular phoneme

with approximately a recognition rate of about 75 % (for the native and non-native models). This information will be used directly by our system in order to evaluate the pronunciation.

4. RECOGNITION APPROACH

The basic pronunciation scoring paradigm previously developed [2,3,4] are based on Hidden Markov Models (HMMs) to generate phonetic segmentations of the student's speech. From these segmentations, machines scores are obtained based on HMMs likelihood, phone duration and a combination of these scores.

We propose here to use directly the hybrid system HMM/MLP already trained on a native and a non-native database in order to evaluate the quality of pronunciation and identify the pronunciation problems.

As in [2], speech is modelled as a sequence of phones HMMs trained with native and non-native speech data. We use a network with alternate pronunciations where each phone can optionally be pronounced either as a native or as a non-native model as show in figure 1.

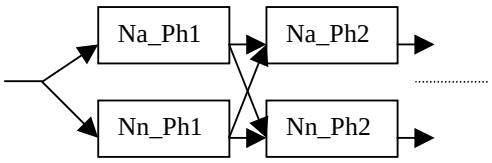

Fig 1 : The Mispronunciation Network (MP) where each phone has two alternative pronunciations : Na for native and Nn for non-native.

The native and non-native hybrid HMM/MLP systems are described in the previous section.

To detect mispronounced phones, we assumed knowledge of the orthographic transcription of the sentence pronounced. An MP network for the sentence is build and the Viterbi phone backtrace procedure produces a sequence of native and non-native phones with their duration and a relative confidence score for each phone.

The pronunciation scoring paradigm uses the segmentations at the phone level and a score for each phone segment based on log-posterior probabilities given by the NN.

This measure has already proven [4] to outperform other ones like log-likelihood HMM based score or segment duration scores.

$$s_j = \frac{1}{d}.\sum_{t=t_i}^{t_i+d-1}\log(p[q_j \mid X_t])$$

Where d is the duration of phone q_i.

The pronunciation scoring consists in giving the score of each phoneme and a global score for the whole utterance. Analysing more precisely the results in table 1, we found that some phones are less recognized than others. So we decided to weight some phone less heavily than other, depending on their importance for mispronunciation.

$$score_1 = \frac{1}{N}.\sum_{j=1}^{N} w_j.s_j$$

Where N is the number of phones in the sentence.

A much simpler measure commonly used is :

$$score_2 = \frac{nbr\_non\_native\_phones}{nbr\_native\_phones}$$

Those score will be used in order to evaluate the pronunciation and to calculate a human-machine correlation at the sentence level. We are currently asking a human expert to listen and verify some of the speakers of the database and note the mispronounced phones. So we will know to what extent our mispronunciation detection algorithm converge with human judgement.

5. THE FINAL APPLICATION

The final application will include a pronunciation error recognition tool combined with a multimedia didactic component, based on existing learning material created at the Modern Language Department (Namur University).

The error recognition tool consists of a single window with :

- the sentence to be pronounced
- the signal that has been recorded
- the segmentation at the phoneme level using the native and non-native models
- a confidence score at the phoneme level which will be used to decide whether the phonemes have been pronounced correctly or not.

We will try many different confidence measures based on posterior probabilities, segment duration and a combination of them.

In addition, we intend to build a graph where each phone can optionally be pronounced either as a native or as a non-native. In order to detect mispronounced phones

we will used the classical Viterbi phone backtrack which will contain a sequence of native and non-native phones.

At this stage of the project, the didactic component will tackle three specific pronunciation items, in a twofold approach : (i) a brief theoretical account of the difficulty in question, and (ii) a set of exercises applied to the difficulty.

6. CONCLUSION

This paper discusses an original approch for the automatic detection and correction of pronunciation errors for foreign language learners. Particular attention has been devoted to the creation and labelling of a sound database in Dutch, pronounced by natives and non-native speakers. The final application is able to identify errors at the phoneme level, which distinguishes it from its competitors, that provide graphic or numeric feed-back. This result has been achieved by using the well-known hybrid HMM/ANN recognition approach, that combines Hidden Markov Models and Artificial Neural Networks. The first validation data are promising, and the mispronunciation detection algorithm is strongly expected to converge with human judgement.

7. REFERENCES

[1] J.M. Boite, F. Bataille, O. Deroo, S. Dupont, V. Fontaine, C. Ris and Laurent Zanoni : " Un logiciel complet pour l'entraînement et la reconnaissance de la parole ", Proc. Premières Journées Scientifiques et Techniques FRANCIL, pp. 41-44, Avignon, 1997.

[2] O. Ronen, L. Neumeyer and H. Franco : « Automatic Detection of Mispronunciation for Language Instruction », Proc. EuroSpeech 97, Rhodes, Greece, pp. 649-652.

[3] Y. Kim, H. Franco, and L. Neumeyer : « Automatic Pronunciation scoring of specific phone segments for Language instructions », Proc. Eurospeech 97, Rhodes, Greece, pp. 645-649.

[4] H. Franco, L. Neumeyer, Y. Kim and O. Ronen : « Automatic Pronunciation scoring for Language Instruction », Proc. ICASSP 97, April 97, Munich, Germany, pp. 1470-1474.

[5] J. Vanparys, G. Deville & S.Gielen : «Démosthène: naar uitspraakremediëring met de computer», ANBF-nieuwsbrief, november1998, pp. 89-102.