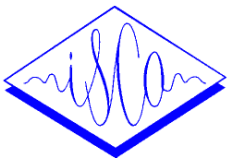

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

AUTOMATIC DETECTION OF MISPRONOUNCED PHONEMES FOR LANGUAGE LEARNING TOOLS

Olivier Deroo[†] *Christophe Ris*[†] *Sofie Gielen*[‡] *Johan Vanparys*[‡]

[†]Faculté Polytechnique de Mons, 31, bld Dolez - B-7000 Mons - Belgique

[‡]Faculté Universitaires Notre-Dame de la Paix, 61, rue de Bruxelles - B-5000 Namur - Belgique

e-mail: {deroo,ris}@tcts.fpms.ac.be | {sofie.gielen,johan.vanparys}@fundp.ac.be

ABSTRACT

Automatic Speech Recognition (ASR) can be very useful in language learning tools in order to correct mistakes in the pronunciation of foreign words by non-native speakers. Most of the systems integrating ASR proposed on the market are just rejecting or accepting whole words or whole sentences. In this paper, we propose a method to identify the pronunciation errors at the phoneme level. Indeed, mistakes are often predictable and concern a particular subset of phonemes not present in the mother language of the speaker. We describe two different approaches based on the Hybrid HMM/ANN technology. The methodology for the training of the recognizer is discussed, and we describe a new approach where a mixed database is used to train a speech recognition system able to detect pronunciation errors at the phoneme level. Preliminary but promising results have been obtained on the DEMOSTHENES database.

1. INTRODUCTION

Acquiring a good pronunciation of spoken sentences in any language is a non-trivial task for most non-native speakers. Traditional audio-visual aids – in classrooms or language laboratories – have shown their limitations in correcting pronunciation (lack of systematic feed-back in a non-individualized environment). On the other hand, recent developments in continuous speech recognition make it possible to provide multimedia tools that analyze and correct the pronunciation of non-native speakers in a consistent and individualized approach. Unfortunately, most of the systems proposed so far on the market are taking binary decisions on whole words or even whole sentences, which basically gives little help on the way to improve one's pronunciation. The system proposed in this paper is able to localize the pronunciation errors at the phoneme level. Such an approach distinguishes the application from commercial products currently available on the market, which provide feedback in a graphic, non-linguistic format. In section 2, we describe the DEMOSTHENES database collected for this particular task. In section 3, we present and discuss the two different approaches we developed. These methods have been tested on this database and are currently extended to other European languages through the L-KIT[1] project.

2. THE DEMOSTHENES DATABASE

This database has been collected in order to train our system and test it over a wide range of speakers in the framework of the DEMOSTHENES project. The goal of this project is to build language-learning tools integrating automatic speech recognition for French speaking people learning Dutch [2, 9].

This database, recorded on microphone, consists in Dutch sentences that are representative of the typical pronunciation errors encountered by the learners (e.g. language-specific phonemes without equivalent in French, assimilations, confusion between long/short vowels, etc.). About 22,000 items have been recorded by 135 native and non-native speakers. Those items have been carefully selected from the basic vocabulary of the Dutch language (covering the 2,000 most frequent words) in order to illustrate the most frequent pronunciation difficulties encountered by French-speaking students.

Basic phonetic units have been labelled in the specific context of DEMOSTHENES, that is an extended phonetic alphabet has been defined for the coding of the speech database, including erroneously pronounced phonemes, so that pronunciation mistakes are labelled as well. The processing of

[1]DEMOSTHENES and L-KIT are research projects sponsored by the Walloon Region of Belgium.

Dutch pronunciation features depends on the mother language of the learners: are considered relevant only the most probable mistakes committed by French speakers learning Dutch. Other mistakes are not labelled as such, and are thus irrelevant in this context (due to their low expectation).

The DEMOSTHENES database has been used in order to test the two methods described in the following sections. The test set consists of 12 native (6 males / 6 females) and 12 non-native (6 males and 6 females) speakers (different from those selected for the training of the speech recognition system) uttering approximately 2,750 sentences. A linguist expert has manually identified and labelled about 2,000 pronunciation errors covering the 11 most important difficulties in Dutch.

3. ASR-BASED APPROACH

The basic pronunciation analysis algorithms proposed in the literature [3, 6, 8] are based on phonetic segmentations of the speech signal automatically generated by forced Viterbi alignment through Hidden Markov Models (HMM) [5]. Given these segmentations, scores are obtained from HMM likelihoods, phone durations or a combination of both of them. These scores can then be used to decide whether the pronunciation is acceptable or not.

In this paper, we propose to use the hybrid system combining Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) trained in a specific classification mode in order to evaluate the quality of pronunciation and precisely identify the pronunciation problems.

In the two approaches introduced in this paper (and as in [8]), speech is modelled by phoneme-based HMMs modelling both the correct and incorrect pronunciations. To detect mispronounced phonemes, we assume that we know the correct orthographic transcription of the sentence pronounced (as it is the case in most language learning exercises where sentences are prompted). A phoneme graph for the sentence is built taking into account both correct and incorrect phoneme models and the most probable sequence of phonemes with their respective duration can be produced by a forced Viterbi alignment [5]. The phoneme graph models in parallel the right pronunciation of each phoneme and the corresponding pronunciation errors. Two different forms of such graphs are used in the methods proposed in this paper.

Moreover a confidence score can be computed for each phoneme in order to help taking decisions on the presence of a mispronounced phoneme. Indeed if the confidence score is too low for a particular phoneme, this could be interpreted as a pronounciation error that has not been properly modelled. This score can be computed from the log-posterior probabilities provided by the ANN (cf. equation 1). This measure has already proven [3] to outperform other ones based on HMM log-likelihoods or segment durations.

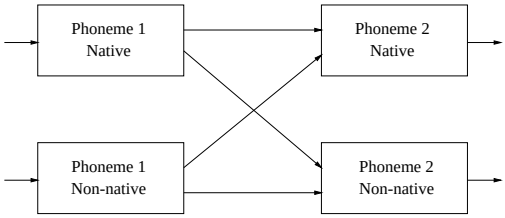

Figure 1: *The phoneme graph where each phone has two possible pronunciations: one for native and another one for non-native.*

$$s_j = \frac{1}{d} \sum_{t=t_j}^{t_j+d-1} \log(p(q_j|X_t)) \quad (1)$$

Where $p(q_j|X_t)$ is the posterior probability of state q_j at time t and d is the duration of phone q_j.

Therefore, if a non-native phoneme is found in the phoneme sequence, we are able to determine precisely the place and the type of the error that has been made.

3.1. Competing models

In the first approach, two hybrid HMM/ANN systems are trained independently on Dutch speech data recorded by native and non-native speakers. Each phoneme of the sentence can optionally be modelled either by a native HMM or by a non-native HMM. The phoneme graph is then composed of a sequel of competing models as show in figure 1 All the ANNs used in the experiments reported in this paper have an input layer of 234 units spanning a window of 9 frames, where each frame consists of 12 cepstral parameters (log RASTA-PLP [4]), their first derivatives, the first and second derivatives of the energy. The log-RASTA-PLP parameters have been chosen because of their robustness against changes in the recording conditions (typically the use of different microphones). As we are working at the phoneme level, we define an output layer of 42 units, corresponding to one unit per Dutch phoneme. The classification accuracy (on both the training set and a cross-

	Train	Cross
Native ANN	80.2%	76.7%
Non-native ANN	76.7%	76.5%

Table 1: *Phone classification rate at the frame level with the ANNs trained on the native and non-native databases. Use of log-RASTA-PLP parameters.*

validation set) obtained at the frame level can be seen in Table 1.

Unfortunately detection of mispronounced phonemes using this approach was not efficient (about 35% of the labelled pronunciation errors were correctly detected). By analysing the behaviour of the system, we noticed that most of the phonemes trained with the native or non-native databases were very closed to each other. Indeed, most of the phonemes of the foreign language are correctly pronounced by the non-native speakers (such as plosives, nasals, ...) so that the system is not able to discriminate between wrong and right pronunciations leading to many false mispronunciation detection.

3.2. Mixed model

The second experiment makes the hypothesis that foreign speakers always make the same kinds of pronunciation errors and that, when they mispronounced a phoneme in a language, they usually use a sound that is commonly used (or similar to one) in their native language. Therefore, the detection of pronunciation errors can be handled the following way. For each sentence prompted to the speaker, the phonetic transcription corresponding to the correct pronunciation is known. Based on linguistic knowledge, it is possible to identify in those sentences the most likely errors at the phoneme level. For instance, in the case of a tool teaching Dutch to French-speaking people, the phoneme **'G'** (like in *gaan*, *dag*) is often mispronounced **'g'**, **'x'** or **'k'**. We therefore build a phoneme graph (see figure 2) taking all these wrong pronunciations into account. Each phoneme is modeled by a HMM for which emission probabilities are estimated by a neural network, trained on both Dutch speech data for estimating the posterior probabilities of the phonemes of the target language and French speech data for estimating the posterior probabilities of the listed potential mistakes. Based on these probabilities, it is possible to find, by Viterbi alignment, the most probable phoneme sequence corresponding to the recorded speech and localize what has been mispronounced.

This method requires to know in advance all the mistakes that could be uttered by the non-native speakers. Practically, only the most probable errors can be taken into account. So to ensure the system to be able to detect unpredicted pronunciation errors, we add a garbage model in parallel as an alternative to listed errors, assuming that if a pronunciation is too far from the good one, it will be detected by this garbage model as an *undefined* error. The garbage model is a particular HMM state which emission probability is computed as the average of the N best probabilities provided by the neural network [1].

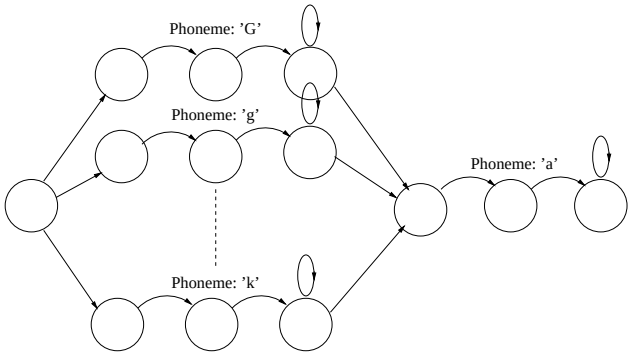

Figure 2: *The phoneme graph where the sentence 'ga zitten' is modelled as a sequence of phonemes pronounced in the right way ('G') and possible pronunciation errors ('g', 'k', ...).*

	Train	Cross
Mixed ANN	77.1%	76.2%

Table 2: *Phoneme classification rate at the frame level with the ANN trained on the mixed database. Use of log RASTA-PLP parameters.*

As we are still working at the phoneme level, we define an output layer of 62 units corresponding to one unit per phonetic class for Dutch: 42 phonemes already defined in the previous section and 20 phonemes from the French language database covering the 11 most frequent pronunciation errors. We used the BREF database [7] in order to train those 20 phonemes extracted from the French language[2]. This model will be called the mixed model in the rest of this paper.

As the ANNs are locally discriminant, we are now able to discriminate between the phonemes that are correctly pronounced or not by the speaker. The classification rates at the frame level for this ANN can be seen in table 2.

We are able to evaluate at each frame, the probability of being in one particular phoneme with an accuracy of about

[2]SAMPA format: e , a , o , y, u, H, S, Z, z, E, e, O, o, 2, 9, @, g, k, s, N

76%. This information will be used directly by our system in order to evaluate the pronunciation.

This system has been evaluated by native and non-native (French) people uttering Dutch sentences. The speakers (12 natives, 12 non-natives) were asked to pronounce several sentences (a total of 2,749 sentences) for which pronunciation errors were manually identified by linguist experts. Eleven different potential difficulties for French speakers have been selected and around 2,000 pronunciation errors have been marked in the test database. The system was able to automatically detect and identify 70% of the manually labeled mistakes.

4. EXTENSION TO OTHER LANGUAGES

The aim of the `L-KIT` project is to build a toolbox that will allow anyone to train specific speech recognition systems to integrate pronunciation error detection in language learning tools and this in as many languages as possible (we are currently working on French-English and French-German). The approach proposed in this paper, of course, needs some strong linguistic knowledge in order to identify as completely as possible the potential pronunciation errors encountered by the speakers. Moreover, the system depends not only on the target language (the teacher) but also on the source language (the student). However, this method proposes an efficient way to introduce ASR in language learning tools. In addition, we plan to incorporate gender dependent models to improve the speech recognition system and also additional tools able to detect the stress in a sentence (using pitch, duration and energy), which is also a source of many pronunciation errors in languages as Italian or Spanish, ...

5. CONCLUSION

This paper discusses an original approach for the automatic detection and correction of pronunciation errors for foreign language learners. Particular attention has been dedicated to the creation and labelling of a speech database in Dutch, pronounced by natives and non-native speakers. The final application is able to identify errors at the phoneme level, with an accuracy of 70%. This result has been achieved by using the hybrid HMM/ANN speech recognition system, that combines Hidden Markov Models and Artificial Neural Networks and by training the system on a mixed database containing the phonemes of the target language and possible sounds from the language of the learner. Exercises can be prepared as series of sentences for which most probable mistakes have been identified. The system can also automatically generates competing phonetic transcriptions of the words in the sentence from a list of predefined pronunciation difficulties. The system is then able to detect the mispronounced phonemes and give back much more accurate advices to improve the pronunciation.

6. REFERENCES

[1] J.-M. Boite, H. Bourlard, B. D'hoore, S. Accaino and J. Vantiegem, *"Task Independent and Dependent Training: Performance and Comparison of HMM and Hybrid HMM/MLP Approaches"*, Proc. ICASSP'94, Adelaide, Australia, Vol.1, pp. 617-620

[2] O. Deroo and G. Deville and H. Leich and S. Gielen and J. Vanparys, *"Automatic Detection and Correction of Pronunciation Errors for Foreign Language Learners : the Demosthenes Application."*, Proc. Eurospeech'99, Budapest, Hungary, Volume 2, pp. 843-846.

[3] H. Franco, L. Neumeyer, Y. Kim and O. Ronen, *"Automatic Pronunciation Scoring for Language Instruction"*, Proc. ICASSP'97, Munich, Germany, pp. 1470-1474.

[4] H. Hermansky and N. Morgan, *"RASTA Processing of Speech"*, IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[5] F. Jelinek, *"Statistical Methods for Speech Recognition"*, The MIT Press.

[6] 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.

[7] L.F. Lamel, J.-L. Gauvain and M. Eskenazi, *"BREF, a Large Vocabulary Spoken Corpus for French"*, Proc. of European Conference on Speech Communication and Technology, 1991, Vol.2, pp. 505-508

[8] O. Ronen, L. Neumeyer and H. Franco, *"Automatic Detection of Mispronunciation for Language Instruction"*, Proc. Eurospeech'97, Rhodes, Greece, pp. 649-652.

[9] J. Vanparys, G. Deville and S. Gielen, *"Demosthenes: narr uitspraakremediëring met de computer"*, ANBF-nieuwsbrief, november 98, pp. 89-102.