AUTOMATIC DETECTION OF FOREIGN ACCENT FOR AUTOMATIC SPEECH RECOGNITION

Katarina BARTKOVA, Denis JOUVET

France Telecom R&D
Lannion, France
Katarina.Bartkova@orange-ftgroup.com

ABSTRACT

Recognition of foreign accented speech remains among the most difficult tasks in automatic speech recognition. It was observed that using models trained on foreign data together with native models improves the recognition for speakers with foreign accent. However such an approach degrades the recognition performances on native speakers. In order to avoid such performance degradation the degree of accent should be detected prior to the recognition process. In this paper an automatic method of detection of the degree of foreign accent is proposed and results are compared with accent labeling carried out by an expert phonetician. This made possible a better targeting of speakers having a heavy foreign accent which allowed using the foreign accent dedicated model when necessary and thus improving recognition performances on non-native speech without major performance degradation on native speakers.

Keywords: Speech recognition, foreign accent

1. INTRODUCTION

Recognition of foreign accented speech is one of the most challenging tasks in automatic speech recognition. While on native speech recognition performance is now acceptable, the performance degradation on foreign accented speech remains high. One of the reasons of the recognition performance degradation observed on foreign accented speech is that the acoustic models are usually trained only on speech with standard native pronunciations. Non-native speech recognition is not properly handled by native speech models, no matter how much dialect data is included in the training [2]. Moreover, differences between foreign accented speech and native speech occur on two levels: at the acoustic level, as foreign speakers can alter the exact pronunciation of a sound, and at the phonological level, as foreign speakers can omit or insert sounds or substitute one sound with other. Therefore in order to deal efficiently with foreign accented speech, speech recognition systems should handle both variants occurring at the acoustic level [8] and variants at the phonological level [3]. It was observed [1] that introducing pronunciation variants into the phonetic description of the words and adding acoustic models adapted on foreign speech data in the modeling process improve foreign speaker's speech recognition, especially when the accent is heavy. However, such complexity increase leads to lower performance on native speakers [6], [7]. In order to avoid performance degradation on native speakers or foreign speakers with slight accent, it would be effective to use specific modeling techniques only for speakers speaking with heavy accent while avoiding such a use for the other speakers. It was shown [4] that foreign accent can be successfully detected, as a classification among 4 different accents was achieved with an accuracy of 81.5% using multiple acoustic and prosodic features. In [9] an accent classification followed by model selection achieved an absolute reduction of error rate ranging from 1% to 1.4% when the degree of accent varied significantly.

The goal of this study is to present a simple method to detect automatically the degree of foreign accent. Once the degree of the foreign accent is detected, then the best appropriate model can be used for recognition. Such an approach should avoid performance degradation on native speakers or speakers having a slight accent and should improve the speech recognition on speakers with heavy accent.

2. BASELINE OVERVIEW & DATA BASE

2.1. Non Native Speech Corpus

The speech corpus used in this study was collected from French, English, German and Spanish speakers pronouncing French words or expressions. Thus this corpus exhibits several types of foreign accents. The data base was

recorded through the telephone network and contained 83 French isolated words and expressions. The corpus was pronounced by 79 native French speakers (from France), 81 Spanish speakers, 171 English speakers (from UK, USA), and 200 German speakers (Germany).

The degree of accent was "hand labeled" by an expert phonetician distinguishing 8 degrees of accent as it is indicated in Table 2.

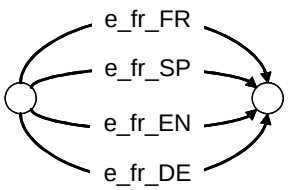
2.2. Speech Modeling

The speech modeling used here is HMM-based, and relies on a context-dependent modeling of the phonemes [5]. The contextual models are defined using a priori knowledge, and for the present study, the contexts were defined in such a way that they were compatible between different languages, making possible a simultaneous use of contextual models from different languages with an adequate handling of contexts at phoneme boundaries. The standard models in each language are trained in a classical way that is, using speech data of the given language and associated pronunciation descriptions.

The acoustic analysis computes MFCC features. Mixtures of Gaussian densities are used for the acoustic modeling of the MFCC and energy features as well as their first and second order temporal derivatives.

For recognition, two acoustic models are used. The first one is a standard native model using native French phoneme models in context and the second one uses native French models together with French models adapted on Spanish or German or English speech data. The data bases of the three languages were transcribed with French phonemes and the French phoneme models were thus adapted on the foreign speech data as described in [1]. The adapted phoneme models were then used in a model topology as illustrated in Figure 1 for the phoneme /e/. A single "foreign adapted" recognizer is used.

Figure 1: Using French native unit and French unit adapted on foreign data for modeling a phoneme



In order to capture the foreign accent on the phonological level, pronunciation variants were automatically introduced. Phonological rules were used to deal with the pronunciation of the vowels having in French open and closed counter-parts (such as /ɛ/ and /e/, /ɔ/ and /o/, /ø/ and /œ/). In fact, in the foreign accent targeted model, at every occurrence of a double aperture vowel both counterparts were systematically allowed. A second set of phonological rules transform the nasal vowels (absent in the three foreign languages studied here) into oral vowel and a nasal consonant and use these sequences as a pronunciation variants together with the nasal vowels. Finally a last set of phonological rules implemented in this study allowed the use of back-rounded vowels or semivowels as pronunciation variants to front-rounded vowels and semivowels (again absent from the phoneme inventory of the three languages). In the following, the model containing adapted phoneme models as well as foreign pronunciation variants will be referred to as "foreign adapted" while the base-line model, containing only French modeling units will be called "native". As reported in Table 1, the model "foreign adapted" worked better for foreign speakers (except for German speakers) while the model "native" yielded better results for French speakers. Good results obtained on German speakers (in Table 1 as well as in Figure 6 later) are probably due to a great overlap between French and German vowel inventories. In fact, German like French has a tense pronunciation, open and closed vowels, rounded front vowel and uvular [r].

Table 1: Speech recognition error rates yielded by the native and the foreign adapted models.

	French	Spanish	German	English
native	5.2%	14.8%	5.2%	13.0%
foreign adapted	8.6%	9.8%	5.5%	8.6%

3. ANALYSIS OF THE SPEECH CORPUS

It seemed important to analyze whether there is a correlation between the hand labeled accent degrees and the automatic speech recognition results. In order to do so, recognition tests were performed using the native model involving only native French modeling units. As it is reported in Table 2, recognition error rates are rather consistent with the expert judgment based labeling, though probably the 8 degree accent scale is too fine. In fact, the system seemed to have increasing

difficulty with mid and heavy accents while there was no substantial difference between processing slight, very slight or no-accent groups.

Table 2: Manual accent annotation and automatic recognition performance

Expert notation	Degree of accent	Number of speakers				Error rate (%)
		FR	EN	DE	ES	
no-accent	0	110	14	57	11	6.7
very slight	1	2	11	26	4	6.3
slight	2	1	23	73	13	6.7
slight mid	3		13	16	7	8.2
mid	4		33	28	16	11.2
heavy mid	5		10		6	18.9
heavy	6	1	52	3	15	30.2
very heavy	7		25		8	38.2

4. ACCENT DETECTION

In order to detect the degree of the foreign accent automatically, a free phoneme loop decoding is carried out using context dependent models of French phonemes together with context dependent models of one of the three languages. Thus, for each utterance, 3 free phoneme loops of decoding are performed (one for each modeling of the phonemes as indicated in Figure 2). The percentage of usage of French phonetic units along each best decoding path is computed, and average over the 3 models.

Fig. 2: Modeling units used for accent detection.

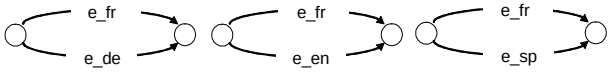
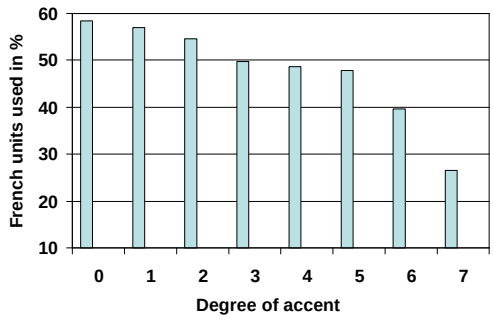


Figure 3 shows these average percentages of usage of French phonetic units along the best decoding paths as function of the manually annotated accent degree of the speech data.

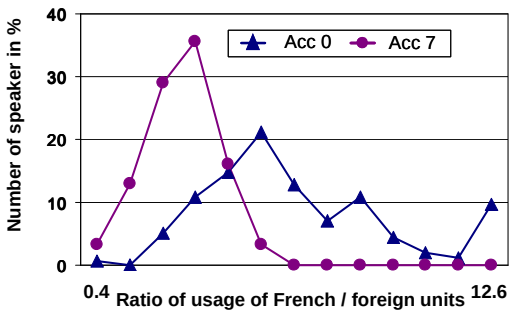
Figure 3: Percentage of French units used as a function of the degree of the accent



When the degree of accent is low (no accent or a slight one, typically French speakers and some foreign speakers), the number of French units used in the alignment with the free phoneme loop model is high. As the degree of accent increases the percentage of usage of French units diminishes. And when the accent is very heavy, then mostly foreign modeling units are used.

Also, the ratio of the number of French units over the number of foreign units, observed along the best matching path, was calculated. No substantial differences were observed among the ratios obtained with the different models (fr//de; fr//en and fr//sp). An average ratio (over the 3 alignments) was then calculated. Histograms of the number of speakers as function of this ratio of usage of French vs. foreign units were computed. As it can be seen on Figure 4, a separation can be found between degrees of accents that are not closely situated in the expert notation (such as heavy accent vs. no-accent).

Figure 4: Histogram of amount of speakers per ratio of usage of French over foreign units for accent degree 0 (no-accent) and degree 7 (heavy accent).



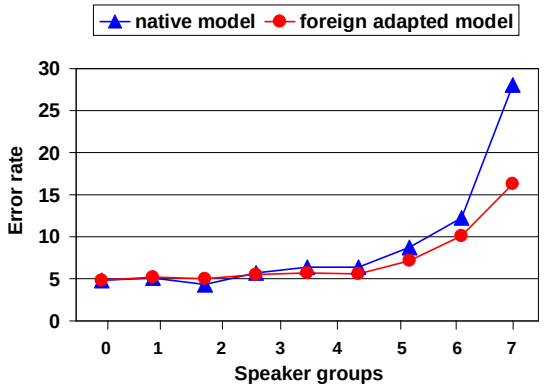
However, such a histogram based separation is not possible for accent degrees that are close, for example no-accent (0) vs. very slight accent (1) or slight accent (2)

5. EVALUATION AND DISCUSSION

In order to test the efficiency of the approach, speakers were grouped according to their ratio of usage of French and foreign units into 8 groups. This grouping can be considered as an automatic recognition of the accent degree: the lower the ratio of usage of the French modeling units is, the stronger the speaker's accent is. Figure 5 shows that the use of the model "foreign adapted" (using adapted phoneme models and containing foreign pronunciation targeted phonological rules) is beneficial especially to the strongest accent (groups 6 & 7 corresponding to the lowest use of French modeling units). In fact, the error rate

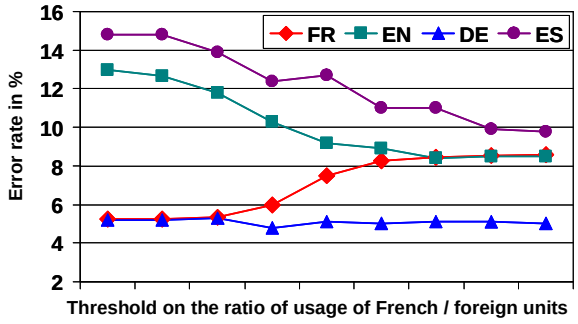
reduction on theses two groups is about 42% (7th group containing 33 speakers) and 18% (6th group containing 80 speakers) respectively.

Figure 5: Error rate for automatically created accent groups according to the ratio of usage of French units



Comparing the ratio of usage of the acoustic models to a threshold would allow using the most appropriate model for each speaker: a native model when there is a high ratio of usage of French units and an adapted foreign accented speech recognition model when the ratio is low. Figure 6 gives an idea of the error rate evolution for the 4 native speaker's groups (French, German, Spanish and English) when the threshold evolves between very high (only native model is used – left hand side) and very low (only "foreign adapted" model is used – right hand side).

Figure 6: Error rates according to the threshold applied on the ratio of usage of French units for selecting the recognition model (native or accented)



It is worthwhile mentioning that expert notation not always reflects speech recognition error rates: speech with an accent considered as heavy by an expert can be perfectly recognized whereas a non-accented pronunciation can be poorly recognized.

6. CONCLUSION

In this study an attempt is made to establish an automatic classification of speakers according to their degree of foreign accent when speaking French. Four speaker groups were used: French, German, Spanish and English. In the accent decision procedure, French acoustic modeling units were used together with standard acoustic units of the 3 foreign languages. The ratio of usage of French units in a free phoneme decoding allows choosing the most appropriate model in the recognition procedure: a native model, involving only French acoustic modeling units for speakers having no-accent or only a slight one (when the free phoneme decoding uses a low amount of foreign units) and a foreign accented speech recognition model (containing specific phonological rules and adapted acoustic models) for speakers considered as having a heavy accent in French. This way of doing reduces performance degradation on native speakers compared to a systematic use of the "foreign adapted" model and avoids poor recognition performance for foreign speakers having a heavy accent compared to a systematic use of the "native" model.

7. REFERENCES

[1] Bartkova, K., Jouvet, D. 2004. Multiple models for improved speech recognition for non-native speakers, *Proc. Specom*, Saint Petersbourg, Russia.

[2] Beattie, V., Edmondson, S., Miller, D., Patel, Y., Talvola, G. 1995. An integrated multidialect speech recognition system with optional speaker adaptation, *Proc. Eurospeech*, 1123-1126.

[3] Bonaventura, P., Gallochio, F., Mari J., Micca, G. 1998. Speech recognition methods for non-native pronunciation variants, *Proc. ISCA Workshop on modelling pronunciation variations for automatic speech recognition*, Rolduc, Netherlands, 17-22.

[4] Hansen H.L.J, Levent M.A 1995, Foreign accent classification using source generator based prosodic features, *Proc. ICASSP*, pp 836-839.

[5] Jouvet, D., Bartkova, K. Monné, J. 1991. On the modelization of allophones in an HMM based speech recognition system, *Proc. Eurospeech*, Genova, 923-926.

[6] Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., Zavaliagkos, E. 1994. Comparative experiments on large vocabulary speech recognition. *Proc. ICASSP*, Volume I, 561-564.

[7] Lawson, A.D., Harris, D.M., Grieco, J.J. 2003. Effect of foreign accent on speech recognition in the NATO N-4 Corpus", *Proc. Eurospeech*.

[8] Witt, S. Young, S. 1999. Off-line acoustic modelling of non-native accents, *Proc. Eurospeech*, Budapest, Hungary, 1367-1370.

[9] Zheng Y., Sproat R. & al, 2005, Accent detection and speech recognition for shanghai-accented mandarin, *Proc. Interspeech*, Lisbon.