USING MULTILINGUAL UNITS FOR IMPROVED MODELING OF PRONUNCIATION VARIANTS

K. Bartkova & D. Jouvet

France Télécom – Division R&D – TECH/SSTP
2 avenue Pierre Marzin, 22307 Lannion, France

ABSTRACT

Standard speech modeling generally implies the combination of models of the phonemes of the current language with a description of possible pronunciation variants of the vocabulary words. When dealing with foreign accent, this standard native speech modeling is not adequate. In fact many variabilities have to be taken into account as the acoustic realization of the sounds by non-native speakers does not always match with native models and some phonemes may be replaced by others. By introducing models of phonemes estimated from speech data of other languages, and adding extra pronunciation variants through phonological rules, speech recognition performance improvements were achieved on non-native speech. In this study, a selection of the most frequently used variants is proposed, which relies on the frequency of usage of the various models associated to each phoneme on a development set. Although this selection process is rather simple it provides significant performance improvement.

1. INTRODUCTION

The recognition of native speech reaches in many cases an acceptable level, however the processing of non-native accent remains among the most difficult tasks [1]. The issue of foreign accent is getting crucial in automatic speech recognition as speech enabled information services will be more and more used by non-native speakers for our society is getting ever more cosmopolitan. When words or sentences are uttered with an altered pronunciation, be it a regional or a foreign accent, speech recognition performance is in many cases considerably lowered. Foreign accent is harder to handle than regional accent, because it is less homogenous and it depends on the influence of the native language and on how capable the speaker is to imitate the pronunciation of the target language. This is due to the fact that speakers speaking a foreign language can slightly alter some phoneme features: for example aspirated stops can become non aspirated. Also, speakers can replace a phoneme of the target language, which is absent in their native language phoneme inventory, with the one they consider as the closest in their native language [2]. Moreover, even for the same native language influence there are several degrees of foreign accent.

The recognition performance degradation observed on foreign accented speech has several causes. One of the reasons is that the acoustic models are generally trained only on speech with standard native pronunciation. Accented speech could be added into the training data base. However the danger is that such too "heterogeneous" pronunciations may not be efficient for any foreign accent albeit the resulting modeling may be less efficient for standard pronunciation. Moreover, differences between foreign accented speech and native speech occur also at the phonological level [3]. Therefore, for dealing efficiently with foreign accented speech, speech recognition systems should handle variants occurring at the acoustic level [4] and at the phonological level [5].

This paper is organized as follows. Section 2 describes the speech corpus used and analyses the results obtained with the baseline system relying on native speech modeling. Section 3 summarizes the introduction of foreign-based units and phonological rules, and discusses the corresponding recognition performances. Section 4 introduces selection processes for optimizing the amount of added variants and analyses the corresponding speech recognition results.

2. BASELINE OVERVIEW

The speech corpus used in this study was collected from speakers originating from 24 countries. The corpus contained French isolated words and expressions collected over the telephone. It was split into two parts, one used as a development set, and the other as test set. In all the reported experiments the recognition vocabulary contained 83 words and expressions. Some phonetic forms were rather similar, as for example "oui" (w.i ⇔ *yes*) and "huit" (Y.i.t ⇔ 8).

For analyzing the speech recognition performances, the test set was divided into subsets, each one corresponding to the data (French utterances) pronounced by speakers originating from a given language group. 11 language groups were defined.

- French group: 94 speakers from France, Belgium, Switzerland, Canada, Guadeloupe, Reunion, ...
- Spanish group: 35 speakers from Spain.
- English group: 96 speakers from USA, UK, Ireland & Australia.
- German group: 113 speakers from Germany & Austria.
- Italian group: 56 speakers from Italy.
- Portuguese group: 17 speakers from Portugal.
- African group: 50 speakers from Senegal, Congo, Mali, … (these countries are all francophone and moreover there was no indication of the speaker mother tongue).
- Arabic group: 53 speakers from Algeria, Tunisia & Morocco.

- Turkish group: 53 speakers from Turkey.
- Cambodian group: 48 speakers from Cambodia (the Cambodian – Khmer – language does not contain tones).
- Asian group: 69 speakers from China & Vietnam. (these two languages are tonal languages).

The speech recognition system is HMM-based and relies on a context-dependent modeling of the phonemes. Mixtures of Gaussian densities were used and applied on Mel frequency Cepstral features, including first and second temporal derivatives.

The corresponding baseline results are available in Figure 2 (leftmost bar). The French speakers obtained the lowest error rate, and large differences were observed in the error rates among the various language groups: from less than 6% for German speakers up to 12% and more for Spanish and English speakers.

3. MODELING PRONUNCIATION VARIANTS

In order to handle non-native speech accents, extra pronunciation variants were introduced in the lexicon descriptions. As non native speakers may pronounce phonemes either as they are pronounced in the target language or as they are pronounced in their mother tongue, pronunciation variants can be defined based on the usage of models of phonemes of the mother tongue. This performs well in a bilingual approach, as described in [3] and [4], when the mother tongue of the speaker is known. But when the speaker mother tongue is not known (e.g. non-native speakers calling a speech enabled service) it is only possible to rely on an enriched modeling using various available foreign models. Another way of enriching the modeling consists in using models of phonemes of the target language adapted on speech data from foreign languages, as detailed in [6]. Finally, phonological rules are useful for generating pronunciation variants ([3], [5] and [6]) by replacing some phonemes of the target language by others (often influenced by native language pronunciation).

3.1. Phone models from foreign languages

This approach consists in allowing for each phoneme, the model of the phoneme in the target language (e.g. e_fr for the French model of the phoneme /e/) as well as the models of the corresponding phonemes in other languages (here the models of /e/ in Spanish, English and German) as represented on the let part of Figure 1.

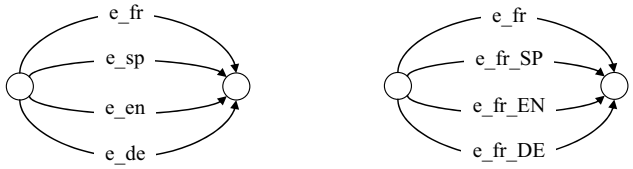

Figure 1 – Adding foreign standard units (on left) or foreign-adapted units (on right) for modeling pronunciation variants.

Standard models of the phonemes in each language are trained in a normal way, using speech data and associated pronunciation descriptions. Context-dependent modeling is defined in such a way that models from different languages can be glued together with proper handling of the contexts.

3.2. Target phone models adapted on foreign speech

In this approach, the added units correspond to models of phonemes of the target language adapted on speech data from foreign languages, as represented on the right part of figure 1, with models of the French phoneme /e/ adapted on Spanish data (e_fr_SP), English data (e_fr_EN) and German data (e_fr_DE).

The correspondences between the phonemes of the target language and the phonemes of each foreign language are handled during the adaptation process. As an example, let's take the words "Paris" and "message" from the English corpus. Their standard English pronunciations are:

Paris_uk ⇔ p_uk . a_uk . r_uk . i_uk . s_uk
message_uk ⇔ m_uk . e_uk . s_uk . ɪ_uk . dʒ_uk

where the suffix "_uk" indicates a British pronunciation, given in terms of British-English phoneme units. In order to match these English units with French units, either simple correspondences between units are used, such as a_uk ⇔ a_fr, or more complex ones, such as dʒ_uk ⇔ d_fr . ʒ_fr. Applying these transformations on every lexicon pronunciation lead to the following descriptions:

Paris_uk ⇔ p_fr . a_fr . r_fr . i_fr . s_fr
message_uk ⇔ m_fr . e_fr . s_fr . ɪ_fr . d_fr . ʒ_fr

These transformed descriptions are then used for adapting the models of the French phonemes on English speech data. The few phonemes that do not have counterpart in the target language can be either associated to some garbage units or the corresponding sentences can be ignored during the adaptation process.

3.3. Phonological rules

The first set of phonological rules deals with vowels having two apertures and no significant timber difference, such as [e, ɛ] and [ø, œ]. Both open and closed vowels variants are then possible:

e ➔ (e + ɛ)

The second set of phonological rules handles possible denasalisation of nasal vowels. A French nasal vowel may be decomposed into the oral vowel which corresponds to the same vocalic timber as the nasal one followed by a nasal consonant which articulation place depends on the following consonant:

ɛ̃ ➔ ɛ̃ + ɛ N

where the phonetic realization of 'N' depends on the right context [n] before apical consonants, [m] before labial consonants and [ŋ] before velar consonants.

The third set of phonological rules handles the possible replacement of the French front round vowel [y] by a back round vowel [u] and of the French front rounded semi-vowel [ɤ] by the back rounded semi-vowel [w]. In fact, speakers having a heavy accent in French often show difficulty uttering front rounded vowel when their native language does not contain such a vowel.

3.4. Results and discussion

In the reported experiments, the target language is French, and the foreign standard units used correspond to Spanish, English and German languages. Foreign-adapted models of the French phonemes have been estimated on speech data from these 3 languages. Results are reported in Figure 2. Phonological rules alone (second bar) significantly improve on some language groups but are useless for others. Introducing foreign standard units (third bar) improves only for Spanish and English speakers, which are two of the languages from which foreign units were added; however no improvement is observed for the other groups. When foreign adapted units are used (fourth bar), without phonological rules, recognition performance degrades.

The last two bars correspond to both the application of phonological rules, and the introduction of either foreign standard

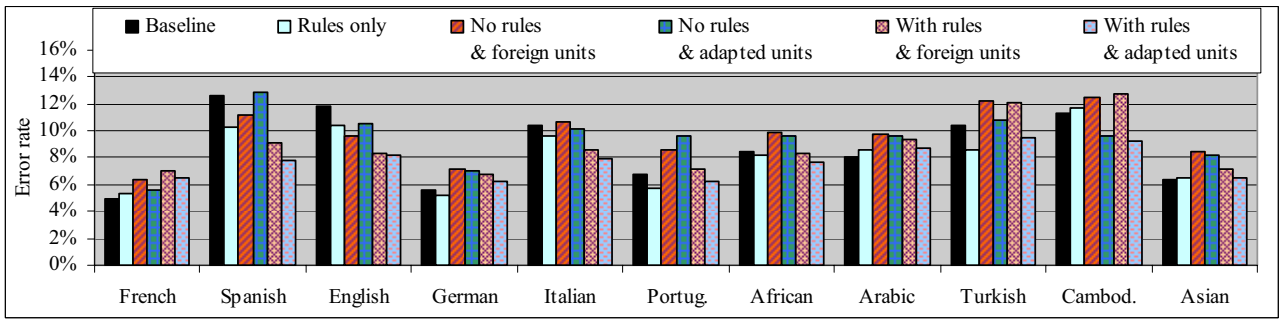

Figure 2 – Error rates for each language group for baseline and modeling variants.

units (fifth bar) or foreign-adapted units (sixth bar). For foreign standard units, a large improvement is observed for Spanish and English speakers, but there is no large improvement with respect to baseline modeling for other language groups. On the other hand when foreign-adapted units are used together with phonological rules, performance improves on many language groups.

4. SELECTING PRONUNCIATION VARIANTS

Adding pronunciation variants in the modeling improves significantly the recognition performance for non-native speakers corresponding to the languages of the added units. However the improvement, if any, is smaller for other language groups and some degradation is observed for French speaking speakers. One can presume that some of the added variants may not be useful, or may even be harmful for some speaker categories. Hence the investigation of selection processes for optimizing the variants.

4.1. Vowel variants only

In [7] an analysis of the frequency of usage of foreign units in the recognition of non-native speech was initiated. It appeared that French units were the most often used for all language groups and the second most frequently used units were generally those corresponding to the native languages of the speakers. Furthermore, it appeared that the usage of foreign standard units was not evenly spread over the phonemes. Usage of foreign units was higher for vowels than for consonants. Therefore, experiments were conducted in which foreign standard units or foreign-adapted units were introduced as variants only for the vowels; the consonants being modeled with the French standard models only.

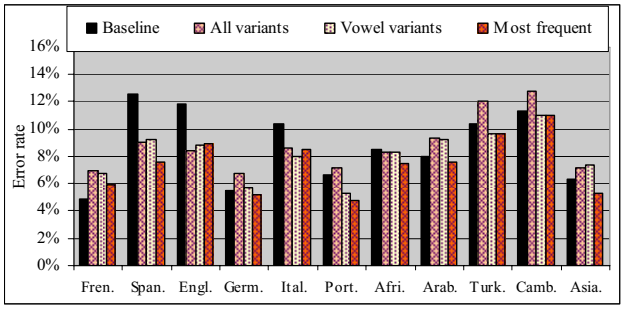

Figure 3 – Error rates for each language group using phonological rules and various sets of foreign standard units.

Comparing second and third bars in Figure 3 shows that limiting the usage of foreign standard units, as variants for the vowel sounds only, provides on average as good results as using foreign standard units as variants for all the phonemes. Averaging the results per category (as displayed in summary table 1), shows that recognition performances are actually better.

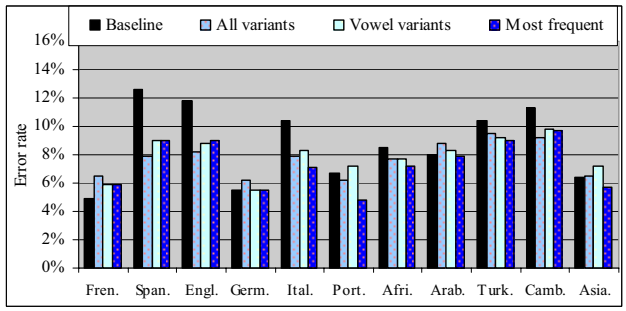

Figure 4 – Error rates for each language group using phonological rules and various sets of foreign-adapted units.

Figure 4 reports similar results for foreign adapted units. Limiting the variants to the vowel sounds (third bars) leads to better results on French speakers, similar recognition performance for speakers from languages used for adaptation (i.e. Spanish, English and German speakers), but on average to slightly worse results for the other language groups.

4.2. Selecting the most frequently used units

In [7], large differences in the frequency of usage among the phonemes were observed. Moreover, across the language groups it was not always the same phoneme that was the most frequently replaced by its corresponding foreign unit. For example the Spanish model for /b/ was frequently used by Spanish speakers speaking French, but the English and German models for /b/ were seldom used for English and German speakers speaking French.

Simple selection processes were used. The basic underlying idea was to align the development set data on the pronunciation variants based either on foreign standard units or on foreign-adapted units. Then for each phoneme, frequencies of usage of the variants were estimated. Two selection processes were experimented. The first one simply kept as variants for each phoneme the most frequently used variant on the development set, or the 2 most frequently used, 3 most frequently used, … and so on. On the opposite, in the second selection process, for each phoneme, only the variants that were used at least x% of the time on the development set were kept. For the results reported in

Figure 5, x took the values 5%, 15% and 25%. The lower this threshold was, the more variants that were used for each phoneme.

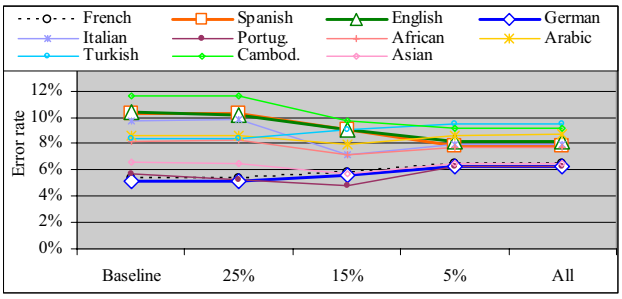

Figure 5 – Error rates for each language group using phonological rules and various selections of foreign-adapted units.

Figure 5 displays baseline results per language group on the left, and results using all the foreign-adapted units for each phoneme on the right. The curves show a smooth behavior with respect to the amount of variants selected. When enough variants are used (e.g. 5% and "all" cases), performances are more homogenous across language groups. An interesting point is that for groups for which performance degrades when all variants are used (for example French and German speakers), the degradation is reduced when only a limited amount of variants are introduced.

As for the languages for which no adapted units are available, Figures 4 and 5 show that by limiting the amount of foreign-adapted variants (here those used more than 15% of the time on the development set), recognition performance improves on many language groups compared to the usage of all available variants.

Table 1 – Summary of results, with various sets of units; phonological rules are also applied when foreign standard or foreign-adapted units are introduced.

	French	Span., Engl. & Germ.	Other lang. groups
Baseline	4.89 % (+/- 1.21 %)	8.95 % (+/- 1.00 %)	8.88 % (+/- 0.85 %)
Baseline + phono. rules	5.38 %	7.86 %	8.56 %
Foreign, all variants	7.01 %	7.70 %	9.44 %
Foreign, vowel var. only	6.76 %	7.37 %	8.66 %
Foreign, selected variants	5.95 %	6.99 %	7.91 %
Adapted, all variants	6.52 %	7.21 %	8.05 %
Adapted, vowel var. only	5.95 %	7.25 %	8.28 %
Adapted, selected variants	**5.87 %**	**7.37 %**	**7.50 %**

The average error rate on those language groups, as displayed in the last column of table 1, drops from 8.88 % for the baseline native modeling, to 8.05 % when all foreign-adapted variants are used, and to 7.50 % when a selection process is applied; which is well out-of the confidence interval indicated with the baseline results. The selection process does not affect a lot the performances on language groups corresponding to the added units, and reduces notably the degradation on the French speakers.

5. CONCLUSION

This paper presented techniques for improving the recognition of non-native speech. Modeling pronunciation variants is necessary for handling non-native speech variabilities. The application of phonological rules helps handling the replacement of some phonemes by others. In complement, introducing models of phonemes in foreign languages provides a way of modeling the realization of sounds by non-native speakers. Such a modeling is efficient when units from languages corresponding to the origin of the non-native speakers can be used. Another approach, based on the adaptation of models of French phonemes on foreign speech data proved to be more effective and more robust, even for speakers from other language groups.

Previous studies analyzing the usage of foreign units on non-native speech led to investigating the reduction of the amount of variants by selecting the most relevant ones. It was observed that foreign models were more frequently used for vowels than for consonants, and experiments showed that, limiting the introduction of variants to vowels only, provided an improvement when foreign standard units were used, but not when foreign-adapted units were used. Finally, simple selection methods were implemented. Reducing the amount of variants that are introduced, by keeping only the variants that were the most frequently used on a development set, led to the best and most homogenous recognition results across various language groups. Optimizing the modeling of the pronunciation variants is thus important, and needs to be investigated further.

6. ACKNOWLEDGEMENT

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-002034 (DIVINES – Diagnostic and Intrinsic Variabilities in Natural Speech). The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

7. REFERENCES

[1] R. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation", *Speech Communication*, vol. 42, pp. 109-123, 2004.

[2] J.E. Flege, C. Schirru, and I.R.A. MacKay, "Interaction between the native and second language phonetic subsystems", *Speech Communication*, vol. 40, pp. 467-491, 2003.

[3] K. Bartkova K. and D. Jouvet, "Language based phone model combination for ASR adaptation to foreign accent", *Proceedings ICPhS'99, International Conference on Phonetic Sciences*, San Francisco, USA, vol. 3, pp. 1725-1728, 1-7 August 1999.

[4] S. Witt and S. Young, "Off-line acoustic modelling of non-native accents", *Proceedings Eurospeech'99, 6th European Conference on Speech Communication and Technology*, Budapest, Hungary, pp. 1367-1370, September 1999.

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

[6] K. Bartkova & D. Jouvet: "Multiple models for improved speech recognition for non-native speakers", *Proceedings SPECOM'2004, 9-th International Conference on Speech and Computer*, St Petersburg, Russia, 20-22 September 2004.

[7] K. Bartkova & D. Jouvet: "Ensemble élargi de phonèmes pour la reconnaissance de parole avec accents", *Proceeding MIDL'2004, workshop on Language & dialectal variety identification by humans & machines*, Paris, France, 29-30 November 2004.