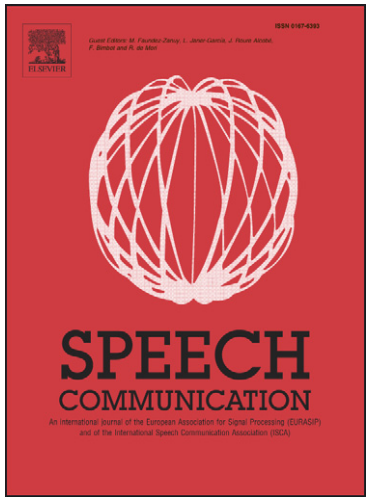
On using units trained on foreign data for improved multiple accent speech recognition

Katarina Bartkova, Denis Jouvet



On using units trained on foreign data for improved multiple accent speech recognition

Katarina Bartkova, Denis Jouvet

PII: S0167-6393(07)00003-9
DOI: 10.1016/j.specom.2006.12.009
Reference: SPECOM 1603

To appear in: *Speech Communication*

Received Date: 1 April 2006
Revised Date: 31 October 2006
Accepted Date: 20 December 2006

Please cite this article as: Bartkova, K., Jouvet, D., On using units trained on foreign data for improved multiple accent speech recognition, *Speech Communication* (2007), doi: 10.1016/j.specom.2006.12.009

On using units trained on foreign data for improved multiple accent speech recognition

Katarina Bartkova & Denis Jouvet

France Télécom – Division R&D / TECH / SSTP
2, avenue Pierre Marzin, 22300 Lannion, France

Corresponding author – Denis Jouvet
E-mail: denis.jouvet@francetelecom.com
Tel: (+33) 2 96 05 34 78

ABSTRACT

Foreign accented speech recognition systems have to deal with the acoustic realization of sounds produced by non-native speakers that does not always match with native speech models. As the standard native speech modeling alone is generally not adequate, it is usually extended with models of phonemes estimated from speech data of foreign languages, and often complemented with extra pronunciation variants. In this paper, the focus is set on the speech recognition of multiple non-native accents. The speech corpus used was recorded from speakers originated from 24 different countries. The introduction of models of phonemes of the target language adapted on foreign speech data is presented and detailed. For the recognition of non-native speech comprising multiple foreign accents, this approach provides better performance than the introduction of standard foreign units. The selection of the most frequent acoustic variants for each phoneme is also discussed as this method makes recognition results more homogenous across speaker language groups. Furthermore, the adaptation of the acoustic models on non-native speech data is studied. Results show that detailed models, which include the modeling of extra pronunciation variants through acoustic units estimated on foreign data, benefit more from the task and accent adaptation process than baseline standard models used for native speech recognition. In addition, experiments show that an adaptation of the acoustic models on a limited set of foreign accents provides speech recognition performance improvements even on foreign accents absent from the adaptation data.

KEYWORDS

Non-native speech.
Speech recognition.
Adaptation.
Multilingual units.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

Efficient handling of foreign accent is getting crucial in automatic speech recognition systems as speech enabled information services are increasingly used by non-native speakers in an ever more cosmopolitan society. However, although the recognition of native speech reaches in many cases an acceptable level, speech recognition performance is considerably lowered when the system has to deal with words or sentences uttered with an altered pronunciation due to a foreign accent. While investigating the variability between speakers through statistical analysis methods, Huang et al. (2001) found that the first two principal components correspond to the gender and accent respectively.

Accented speech is associated to a shift within the feature space of phonemes (Van Compernolle, 2001). For native accents, the shift, applied by large groups of speakers, is global and more or less important, but on overall, the acoustic confusability is not changed significantly. In many cases, most of the regional variants of a language are handled in a blind way through a global training of the speech recognition system using speech data that cover all regional variants. Accent classification has been studied since many years (Arslan and Hansen, 1996) based either on phone models (Kumpf and King, 1996, Teixeira et al., 1996) or specific acoustic features (Fung and Liu, 1999). Also, good classification results between regional accents are reported in (Draxler and Burger, 1997) for human listeners on German SpeechDat data, and in (Lin and Simske, 2004) for automatic classification between American and British accents. To handle variants occurring in regional accents, enriched modeling is often used through multiple acoustic models associated to large groups of speakers as in (Beattie et al., 1995, Van Compernolle et al., 1991), or through the introduction of detailed pronunciation variants at the phonetic level (Adda-Decker and Lamel, 1998, Humphries et al., 1996). However, adding too many systematic pronunciation variants may be harmful (Strik and Cucchiarini, 1999) and experiments showed that it was preferable to have models only for a small number of large speaker populations than for many small groups (Beattie et al., 1995, Van Compernolle et al., 1991).

Compared to native speech recognition, performance degrades when recognizing accented speech and even more when recognizing non-native speech (Kubala et al., 1994, Lawson et al., 2003). Foreign accent is harder to handle than regional accent and its processing remains among the most difficult speech recognition tasks (Goronzy et al., 2004). Non-native accent is less homogenous than native accent as it is influenced both by the native language of the speaker and by the level of his proficiency, that is, his capability to imitate the pronunciation of the target language. As the foreign accent shift from the standard pronunciation depends on theses two factors, non-native speech variability is important.

One of the reasons of the degradation of the recognition performance 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 accented data is included in the training (Beattie et al., 1995). This is due to the fact that non-native speakers can replace an unfamiliar phoneme in the target language, which is absent in their native language phoneme inventory, with the one considered as the closest in their native language phoneme set (Flege et al., 2003). In addition,

differences between foreign accented speech and native speech occur also at the phonological level when a phoneme of the target language is replaced with a different phoneme of the same language (Bartkova and Jouvet, 1999). The replacement of some sounds by others, as well as their insertion or deletion, cannot be handled by the usual triphone-based modeling (Jurafsky et al., 2001). Therefore, for dealing efficiently with foreign accented speech, speech recognition systems should handle variants occurring both at the acoustic level (Witt and Young, 1999) and at the phonological level (Bonaventura et al., 1998). Introducing multiple phonetic transcriptions that handle alterations produced by non-native speakers is a usual approach, and is generally associated to a combination of phone models of the native language with phone models of the target language (Bartkova and Jouvet, 1999, Bonaventura et al., 1998, Witt and Young, 1999). When a single foreign accent is handled, some accented data can be used for training or adapting the acoustic models (Aalburg and Hoege, 2004, He and Zhao, 2003, Liu and Fung, 2000, Uebler and Boros, 1999). Alteration rules can be defined from phonetic knowledge or estimated from some accented data (Livescu and Glass, 2000), or inferred using only native speech of both languages (Goronzy et al., 2004). Raux (2004) investigates the adaptation of the lexicon according to preferred phonetic variants. Only adding accented speech into the training data base may not solve the problem, as modeling together too "heterogeneous" pronunciations may not be efficient for any foreign accent, while the resulting modeling may be less efficient for standard pronunciation. It was found that a separate modeling of foreign accents is needed to capture accurately non-native pronunciations (Beattie et al., 1995). When dealing with various foreign accents, phone models of several languages can be used simultaneously with the phone models of the target language (Bartkova and Jouvet, 2004a), multilingual units can be used (Uebler and Boros, 1999) or specialized models for different speaker groups can be elaborated (Cincarek et al., 2004).

In this paper, the focus is set on the recognition of non-native speech comprising multiple accents. The application context corresponds to speech-activated interactive vocal services, where the identity of the speaker, thus its non-native accent, is not known. The paper is organized as follows. Section 2 describes the speech corpus comprising multiple non-native accents as well as the baseline performance obtained with standard native speech models. Section 3 details and discusses the introduction of modeling variants. Several types of variants are considered. Phonological rules and introduction of standard foreign units are first recalled. Then the introduction of models of phonemes of the target language adapted on foreign data is detailed. This section ends by a presentation and a discussion of the recognition results. Section 4 investigates the selection of variants. Finally section 5 analyses the behavior of these approaches when the acoustic models are adapted on non-native speech data. In addition, the adaptation of the models on non-native speech using various subsets of foreign accents is discussed.

2 SPEECH DATA AND BASELINE OVERVIEW

The speech corpus used in this study was collected from speakers originating from various countries and pronouncing French words or expressions. Thus, this corpus exhibits several types of foreign accents.

2.1 Speech Modeling

The speech modeling used in this study is HMM-based, and relies on a context-dependent modeling of the phonemes (Jouvet et al., 1991). This modeling approach of context-dependent units is similar to the triphone approach, but the sharing of the contextual parameters (i.e. output densities associated to entry and exit states of the phoneme models) is defined beforehand according to phonological knowledge. Hence left and right contexts are defined as classes of phonemes sharing same feature sets (place of articulation, voiced/unvoiced, plosive/fricative/nasal/...) and having a similar influence on the acoustic realization of the sound. About 20 phoneme classes are used for specifying the left and right contexts. Being defined according to a priori knowledge, the contexts are compatible among different languages, thus making possible a simultaneous use of contextual models from different languages with an adequate handling of contexts at phoneme boundaries.

The acoustic analysis computes MFCC features and a frame synchronous line adaptation process is also applied which plays a similar role as the mean cepstral subtraction (C. Mokbel et al., 1996). Mixtures of Gaussian densities are used for the modeling of the acoustic features: MFCC and energy, plus their first and second order temporal derivatives.

In all the reported experiments the recognition vocabulary consisted of 83 French words and expressions. These are typical commands for vocal interactive services, such as the digits, the names of days and months and command words such as "*aide*" (help), "*suivant*" (next), "*précédent*" (previous), etc. A few expressions are also present in the vocabulary, such as "*mode d'emploi*" (instructions for use), "*après-demain*" (the day after tomorrow), etc. Some vocabulary entries had rather similar phonetic forms, as for example /m.w.a/ for the word "*mois*" (month) and /m.w./ for the word "*moins*" (less). All the words are common usual words, and many of them contain only one or two syllables.

2.2 Non-Native Speech Corpus

The speech corpus contains isolated utterances of those 83 French words and expressions collected over the telephone. It was recorded from speakers originating from 24 countries. Each speaker uttered the items in a random order according to vocal prompts. The corpus was split into two parts, one part used as an adaptation set (in section 5), and the other as a test set. The types of non-native accents of the test set are detailed below. The adaptation set has a similar size; it exhibits the same types of accents, but was collected from different speakers. This adaptation set was also used for selecting modeling variants in section 4.

For analyzing the speech recognition performance with respect to the accents, the test set was divided into subsets, each one corresponding to the 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, Cameroon[1], …
- Arabic group: 53 speakers from Algeria, Tunisia & Morocco.
- Turkish group: 53 speakers from Turkey.
- Cambodian group: 48 speakers from Cambodia.
- Asian group: 69 speakers from China & Vietnam[2].

Eventually two language group clusters were also defined. The first one named "EsEnDe" corresponds to the merge of the Spanish, English and German groups. This grouping was driven by the availability of trained standard foreign units for these three languages. The second one named "Other" corresponds to the merge of the remaining groups of foreign speakers: Italian, Portuguese, African, Arabic, Turkish, Cambodian and Asian groups.

2.3 Baseline Performance

The baseline recognition results are reported in Figure 3 (native model M1.A1 – leftmost bar). They were obtained with standard context-dependent French acoustic models trained on a large corpus of telephone French speech data (about 190 hours of signal, corresponding to words and sentences collected from several thousands speakers) and standard native pronunciation descriptions of the French vocabulary words. In the figures, the names Mn.Ak refer to both the type of modeling (M1 for baseline, M2 with phonological rules, and so on) and the data on which the acoustic models are trained (A1 for standard training of generic units, A2 to A5 for task adaptation with various subsets of non-native data). Table 1 summarizes the notations used. In Figure 3 and similar figures, the left part reports the word error rates on each cluster, namely French speakers, then EsEnDe cluster (that is Spanish, English and German speakers), and finally for the other speakers, that is speakers belonging to the 7 other language groups. The right part reports the error rates for each language group.

As expected, the French speakers – i.e. native speakers – obtain the lowest error rate; however large differences are observed in the error rates among the various language groups ranging from less than 6% for German speakers up to 12% and more for Spanish and English speakers. The reason of the good recognition performance obtained on German speakers can be partly explained by the fact that the German language as well as the French language have a tense way of pronunciation (Delattre, 1966), and that they share closed and open vowels as well as rounded front vowels (absent from Spanish and English) and the velar [✕ .

[1] The reason for this group is the following: all these countries are francophone; however there was no indication of the speakers' mother tongue.

[2] The reason of maintaining two separate groups between Cambodian and Asian stems from the fact that Chinese and Vietnamese languages are tonal ones while the Cambodian Khmer language is not.

3 ADDING MODELING VARIANTS

In order to handle non-native speech accents, extra pronunciation variants were introduced in the lexicon descriptions. As non native speakers may utter phonemes either as they are pronounced in the target language or as they are pronounced in their mother tongue, pronunciation variants can be defined using models of phonemes of the mother tongue. This performs well in a bilingual approach, as described in (Bartkova and Jouvet, 1999) and (Witt and Young, 1999), when the mother tongue of the speaker is known. However, 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 presented in (Bartkova and Jouvet, 2004a). Furthermore, phonological rules are useful for generating pronunciation variants (Bartkova and Jouvet, 1999, 2004a, Bonaventura et al., 1998) by replacing some phonemes of the target language by others (often influenced by native language pronunciation).

3.1 Phonological Rules

Several sets of phonological rules were defined and used to generate pronunciation variants for the vocabulary words.

The first set of rules deals with vowels having two apertures and no significant timber differences. The French vowels targeted by these rules are [ ] and [ ]. Once these rules applied, no difference is made between open and closed vowels, that is, at every occurrence of a closed vowel its open counterpart is also added into the word phonetic description (and vice versa). For example the non-native pronunciation for the word "*aide*" (help) will be / (+) / instead of the standard native pronunciation / /. The two aperture vowel processing is achieved through the application of 3 pairs of rules (one pair for each of the vowels having two apertures, as indicated before), like the following rules:

→ (+)
→ (+)

In fact, it is supposed that a foreigner, whose native language does not contain a closed / open vocalic opposition, or at least not in the same phonotactic distribution as in French, would have difficulty to make the difference between them. Thus, maintaining systematically the two vowels should allow the speaker to use any of them.

The second set of rules deals with denasalizing nasal vowels, that is, a nasal vowel can be decomposed into the oral vowel which corresponds to the same vocalic timber as the nasal one (no matter which letter occurs in the spelling form) followed by a nasal consonant which articulation place depends on the adjacent consonant: [ ] before apical consonants, [ ] before labial consonants and [ ] before velar consonants. Therefore, for example, the digit "*cinq*" (five) will have the non-native pronunciation / (+) / instead of its standard native phonetic form which is / /. The denasalization is achieved through the application of a set of 4 rules (one for each French nasal vowel), like the following:

→ + N

where the phonetic realization of 'N' is [●], [ɐ̃] or [ɔ̃] depending on the right context (following consonant).

The reason to these transformations is the following: nasal vowels are seldom present in the inventory of the world languages (Valée et al., 1990) and most of the speakers are unfamiliar with the correct oral plus nasal channel coupling. It is worth mentioning that nasal vowel decomposition exists even in French south dialect where the nasal vowel is decomposed into oral vowel plus a nasal appendix.

The third and last set of rules contains two rules to deal with the French front round vowel and semi-vowel. According to this rule, the front rounded vowel [ø] can be replaced by the back rounded vowel [o] and the French front rounded semi-vowel [ɥ] can be replaced 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. These speakers (for example of Spanish, Italian or English origin) make such replacement especially when the same or a very close phonetic form of a word already exists in their own language and when it contains the same phoneme shift (back rounded versus front rounded).

Thus, for example, once all the above phonological rules are applied, the non-native pronunciation description of the French word "*continuer*" (to follow) becomes /k(ɔ̃+●)ɛ̃●(ɥ+ w)(ʒ+ɑ̃)/ whereas the standard native form is /kɔ̃ɛ̃●ɥɑ̃/.

3.2 Phone Models from Foreign Languages

The acoustic modeling was enriched using standard acoustic models of phonemes of a few foreign languages. These standard models of the phonemes in each language were trained in a classical way that is, using large corpora of telephone speech data of the given language and associated pronunciation descriptions (about 60 hours of signal for German data, 80 hours for Spanish data and 140 hours for English data; each corpus – SpeechDat or similar type of corpus – contains words and sentences pronounced by many speakers). For example, English models of the English phonemes were trained using English speech data pronounced by English speakers, as indicated on the left part of Figure 2.

Then, for each phoneme several models were used: on the one hand, the model of the phoneme in the target language (for example ɑ̃_fr_FR for the model of the French phoneme /ɑ̃/ estimated on French speech data), and on the other hand the models of the corresponding phonemes in other languages. In the reported experiments, the foreign standard units used correspond to the Spanish, English and German languages as represented on the left part of Figure 1.

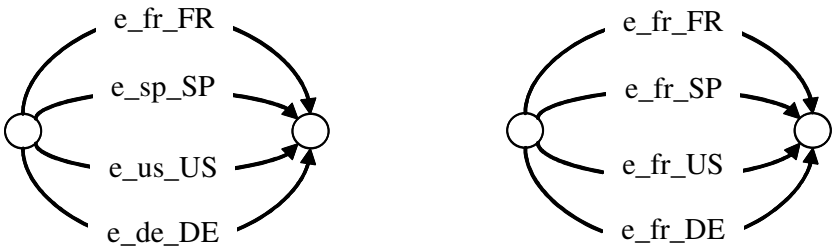


Figure 1 – Adding foreign standard units (left)
or French units adapted on foreign data (right).

In this approach, multiple acoustic models are used for each phoneme, each of them coming from a different language. Hence, it is important to recall that the context-dependent modeling of the phonemes rely on an a priori sharing of contextual parameters according to phonological knowledge. As mentioned before in section 2.1, because of such phonologically-based knowledge, the context-dependent models from different languages can be glued together with a proper handling of the context dependencies.

Having multiple acoustic models of different languages in parallel for each phoneme enables the Viterbi-based decoder to choose for each phoneme the acoustic model that provides the best match according to the input signal, reflecting the degree of foreign accent of the speaker. There are no constraints whatsoever to use units of the same category all along the decoding path (i.e. complete path with native units only or with foreign units only) since such constraints would not be relevant as non-native accent affects phonemes independently of each other, in a speaker and language dependent way.

3.3 Target Phone Models Adapted on Foreign Data

Another way of enriching the modeling consists in using models of phonemes of the target language adapted on speech data from foreign languages (right part of Figure 2). It is expected that such acoustic models will exhibit, at least to some extend, the pronunciation of the target phonemes by non-native speakers.

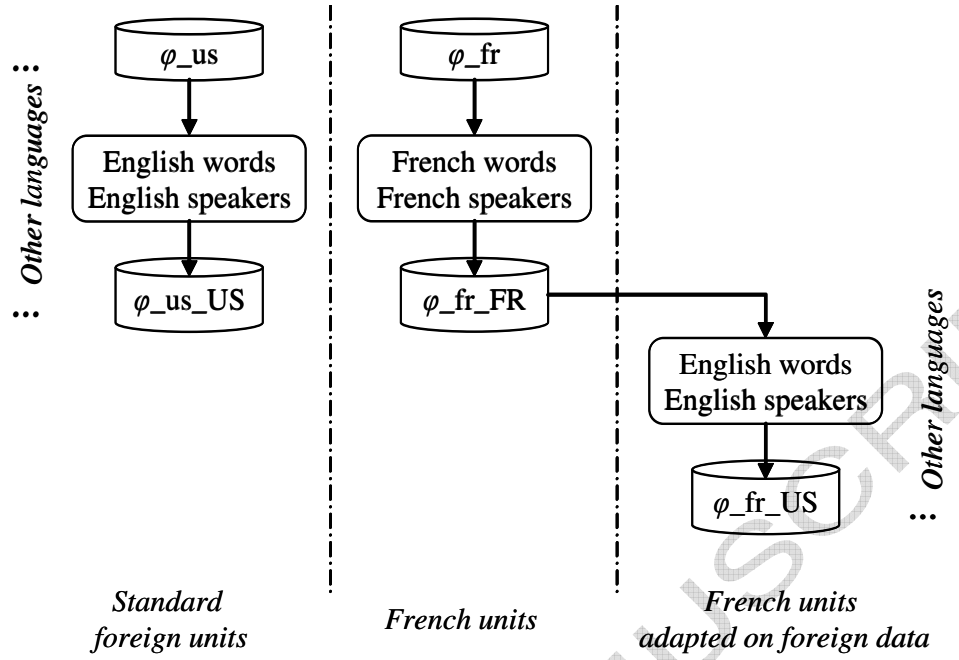


Figure 2 – Training scheme for standard foreign units (left) and French units adapted on foreign data (right).

Through the application of adequate correspondences between the phoneme sets of the foreign language and of the target language, foreign speech data was used to adapt the acoustic models of the target language. The following example explains how French phoneme models were adapted on English data. As an example, let's take the words "Paris" and "message" from the English corpus. Their standard English pronunciations are:

Paris_US ⇔ _us . _us . _us . _us . _us

Message_US ⇔ _us . _us . _us . _us . _us

The suffix '_US' indicates that the utterances belong to the English corpus, and '_us' indicates that the English pronunciations of the words are given in term of American-English phoneme units. In order to match these English units on the French phoneme units, either a simple correspondence between the phoneme units was used, such as:

_us ⇔ _fr

or a more complex one, as for example:

_us ⇔ _fr . _fr

Applying these transformations on every lexicon pronunciation led to the following pronunciation descriptions of these two words:

Paris_US ⇔ _fr . _fr . _fr . _fr . _fr

Message_US ⇔ _fr . _fr . _fr . _fr . _fr . _fr

These transformed pronunciation descriptions were then used for adapting the acoustic models of the French phonemes on English speech data. The acoustic models were adapted using an incremental enrollment procedure which is equivalent to segmental MAP adaptation

with specific choice of priors (Mokbel and Collin, 1999). In the following, such foreign speech adapted units will be denoted as φ_fr_US, for example, where the first suffix, '_fr', refers to the language of the phoneme (here French, that is the target language) and the second suffix, '_US', specifies the origin of the speech data used for adapting the model parameters. The few phonemes which do not have similar counterparts in the target language can be either associated to some garbage units, or more simply, the related sentences can be ignored during the adaptation process.

Subsequently, for the recognition process, each phoneme was modeled simultaneously by the acoustic model corresponding to the target language and a few acoustic models resulting from the adaptation of the target phoneme models on several foreign languages, as represented on the right part of Figure 1. In the reported experiments, the adapted models introduced were the models of the French phonemes adapted on Spanish data (for example ■_fr_SP), English data (■_fr_US) and German data (■_fr_DE): these are the same 3 languages as before.

3.4 Experiments and Discussions

Figure 3 reports the recognition results obtained with the various modeling approaches described before: baseline native acoustic and native pronunciation modeling (M1), baseline acoustic modeling and pronunciation lexicon enriched through the application of phonological rules (M2), then with inclusion of standard foreign units (M3) or French units adapted on foreign data (M4). The last model (M5) will be discussed later in chapter 4. In all cases, generic acoustic models (A1) were used (i.e. these acoustic models are not specific to the recognition task under consideration). The notations used are summarized in Table 1.

Table 1 – Summary of notations for pronunciation modeling and acoustic parameters

Pronunciation modeling	
M1	Native pronunciations only
M2	M1 (native pronunciations) plus variants derived through phonological rules
M3	M2 plus variants associated with standard foreign units
M4	M2 plus variants associated with French units adapted on foreign data
M5	M4, but only most frequent variants are kept for each phoneme

Acoustic parameters	
A1	Generic acoustic models
A2	Task adaptation using data from French speakers only
A3	Task adaptation using data from Spanish, English & German speakers
A4	Task adaptation using data from speakers corresponding to other languages
A5	Task adaptation using data from speakers corresponding to all types of accent

When phonological rules were applied (model M2, second bar) significant recognition performance improvements were observed for some speaker language groups, such as for Spanish, English, Italian, Portuguese and Turkish speakers. The performance improvement can be explained partly by the fact that the first four languages (Spanish, English, Italian and

Portuguese) share with French a great amount of lexical units, having similar spelling forms but with some pronunciation variations. Some of these pronunciation variations were taken into account precisely through the application of the phonologic rules, hence the improvement of the recognition performance for these speaker language groups. However, the approach seemed ineffective for several other language groups. Nevertheless, the introduction of phonological rules improved the overall recognition performance.

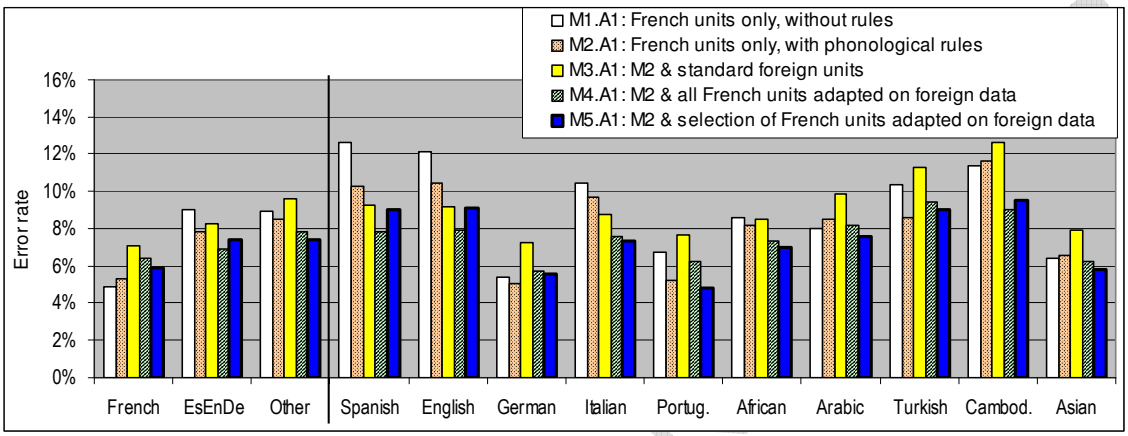


Figure 3 – Error rates for each language group for baseline and modeling variants, before task and accent adaptation.

Introducing foreign standard units in the modeling, besides the phonological rules (model M3, third bar), improved the speech recognition performance for Spanish and English speakers, that is, two of the languages corresponding to the added foreign units; however no improvement was observed for the other language groups, except for Italian speakers. Though German modeling units were also added into the recognition procedure, degradation was observed for this speaker group. One possible explanation of such a less satisfactory performance can be the following: French and German (unlike French and English or French and Spanish) have very similar vowel inventories. Therefore the French modeling units are rather adequate for the recognition of accented speech produced by German speakers. Adding foreign modeling units of other languages augments the perplexity in the decoding procedure and therefore increases the probability of errors as well. In fact, the recognition performance obtained on German speakers evolves in a rather similar way as the performance obtained on French speakers (see Figure 3).

The results obtained with model M4 (fourth bar) show that recognition performance improves on many language groups when acoustic models of the target language phonemes (here French phonemes) adapted on foreign data are used together with phonological rules. Comparing to baseline recognition performance, large error rate reductions are observed on many speaker language groups while a small degradation is observed only for German speakers.

Experiments were also conducted with modeling variants corresponding to the introduction of foreign standard units or the introduction of models of French phonemes adapted on foreign data without the application of phonological rules. Results reported in (Bartkova & Jouvet,

2006) showed small improvements for a few speaker language groups only. Hence this clearly demonstrated that recognition performance improvements on non-native speech result from a simultaneous application of the phonological rules and the introduction of units trained on foreign data.

4 SELECTING MODELING VARIANTS

Adding pronunciation variants, through the introduction of units trained on foreign data in the modeling, improved significantly the recognition performance for non-native speakers corresponding to the languages of the added units. However the improvement was smaller, or even inexistent, for some other language groups and some degradation was observed for French speaking speakers. One can presume that some of the added variants are not useful, or may even be harmful for some speaker categories. Hence the investigation of selection processes for optimizing the variants.

4.1 Preliminary Experiments

In (Bartkova and Jouvet, 2004b), an analysis of the frequency of usage of foreign units in the modeling of non-native French speech was initiated. This was achieved through a forced alignment of the speech signal with an enriched modeling as the one represented in Figure 1, left part. In this model each phoneme of the target language pronunciation description was represented by a plurality of acoustic units associated to different languages. It appeared that French units were the most often used by all speakers of all language groups and that the second most frequently used units were generally those, corresponding to the native language of the speakers.

It was also observed that the usage of foreign standard units was not evenly spread over the phonemes. It 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 while the consonants were modeled with the French standard acoustic models only. Some preliminary experiments were reported in (Bartkova and Jouvet, 2006). Limiting the introduction of variants to the vowels led to recognition performance improvements for some language groups, including for native French speakers, but performance degradations were observed on some other groups.

A detailed analysis of the usage of the variants showed that, across the speaker language groups, not always the same phoneme was the most frequently replaced by its foreign counterpart. For example, the Spanish model for /b/ was frequently used by Spanish speakers speaking French, while the English and German models for /b/ were seldom used by English and German speakers speaking French. This finding led to investigating a phoneme dependent selection of the relevant variants.

4.2 Most Frequently Used

As unrelevant variants may be harmful for the recognition process, a method was elaborated for selecting for each phoneme the most relevant variants. This was investigated in a multiple acoustic modeling framework, using acoustic models of the French phonemes adapted on foreign data. In order to determine which variants are used, the utterances of a development

set are aligned (Viterbi forced alignment) on the non-native pronunciation descriptions which include all variants: application of phonological rules and acoustic modeling variants as in the right part of Figure 1. Then, for each phoneme, frequencies of usage of the acoustic variants were estimated.

Two selection processes were experimented. In the first selection process, only the n-most frequent variants, observed on the development set, were kept with n equal 1, 2, 3, and so on. Whereas, in the second selection process, only the acoustic variants that were used at least x% of the time on the development set were kept. As the selection is applied separately for each phoneme, this second approach leads to a number of variants that may vary from one phoneme to another. For the results reported in Figure 4, x took the values 5%, 15% and 25%. The lower this threshold was the higher was the number of variants used for each phoneme. Only results obtained with the second selection process are detailed below, however, results obtained with the first proposed selection process were quite similar.

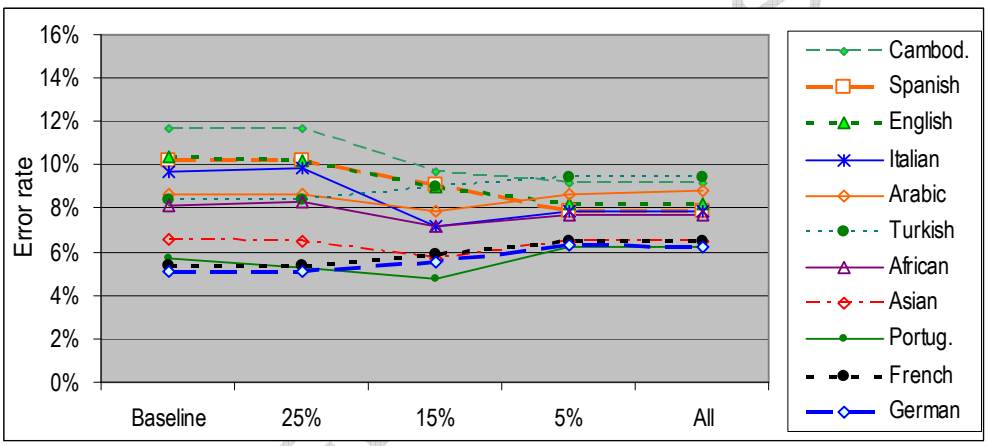


Figure 4 – Error rates for each language group using phonological rules and various selections of foreign-adapted units.

Figure 4 displays recognition error rates per speaker's language group. Baseline results are reported on the left, and results using phonological rules and all the foreign-adapted acoustic units, as well as acoustic variants for each phoneme, are reported on the right. Intermediate results correspond to the selection of different amounts of the most frequently used acoustic variants. The curves show a smooth evolution with respect to the amount of variants selected. When the amount of acoustic variants increases (e.g. from point 25% up to "All"), performance is getting more homogenous across language groups. Results corresponding to the 15% threshold are also reported in Figure 3 (model M5, last bar).

An interesting point is that for groups for which performance degraded when all variants were used (for example French and German speakers), the degradation was less important when only a limited amount of variants were introduced (e.g. point 15%). For Spanish and English speakers, the performance improvements were reduced when the global amount of variants got limited. As for the languages for which no adapted units were available, Figure 4 shows that by limiting the amount of foreign-adapted variants (here those used more than 15% of the

time on the development set), recognition performance improved on several language groups (e.g. Portuguese, Italian, African, Asian, …) compared to the usage of all available variants.

As summarized in left part of Figure 3 and in Table 2[3], with such a selection, the recognition performance degradation on the French speakers was reduced, as compared to the usage of all acoustic variants. Though improvement on the language groups corresponding to the added units was a slightly smaller, better results were obtained on the other language groups.

Table 2 – Summary of results (error rates) with generic acoustic models.

	French	Span., Engl. & Germ.	Other language groups
	1227 utt.	3106 utt.	4322 utt.
M1.A1: Baseline	4.9 % (+/- 1.2 %)	9.0 % (+/- 1.0 %)	8.9 % (+/- 0.9 %)
M4.A1: phonological rules & all fr adapted units	6.4 %	6.9 %	7.8 %
M5.A1: phonological rules & selected fr adapted units	**5.9 %**	**7.4 %**	**7.5 %**

5 ADAPTATION ON NON-NATIVE SPEECH

This section investigates the behavior of the various modeling approaches described before, when the acoustic models are adapted on non-native speech data. Let us recall that the adaptation set had roughly the same size as the test set and that the distribution of the speakers among the language groups was very similar to the one described for the test set in section 2.2.

5.1 All Accents

In a first set of experiments, the acoustic models were adapted using all the data in the adaptation set which exhibited similar non-native accents as those present in the test set. The resulting models were thus adapted to the task environment and non-native accents conditions. The training scheme is represented in Figure 5. The top of the figure refers to the models of phonemes (French, English, Spanish and German languages) that were previously estimated on French or foreign data as described in section 3. The second line recalls the modeling variants (M1 to M5) that were considered, and the arrows indicate the units involved in each case. Every model (M1 to M5) was then adapted to the multiple foreign accent and task conditions using the adaptation set which contains the pronunciation of French words by French and foreign speakers. The acoustic parameters were adapted using an incremental enrollment procedure which is equivalent to segmental MAP adaptation with specific choice

[3] In Table 2, as well as in following tables, the 95% confidence interval is given only for the reference results corresponding to the first line; here the results obtained with the baseline models. This avoids overloading the tables with too many figures.

of priors (Mokbel and Collin, 1999). It should be noted that for each utterance of the adaptation set, the Viterbi alignment selects the most adequate (best matching) variant among all the possible variants of the current model.

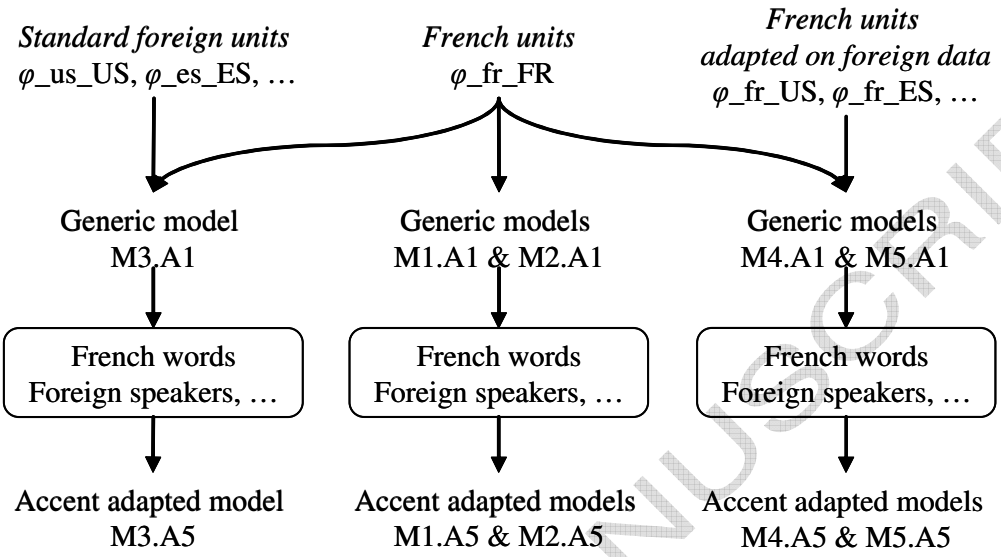


Figure 5 – Task and accent adaptation scheme.

The recognition results obtained with the task and accent adapted models are reported in Figure 6. Obviously, the adaptation improves the recognition performance for all types of modeling. Even after task and all accent adaptation, when the baseline modeling, which consists of only French units with no application of any phonological rules, is used (model M1.A5) the error rates for non-native speakers remains much higher than for French native speakers.

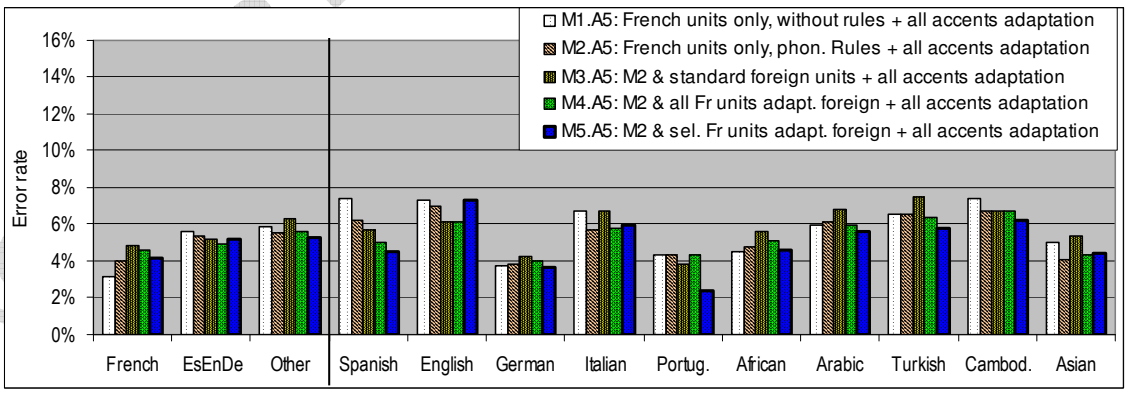


Figure 6 – Error rates for each language group for baseline and modeling variants, after task and accent adaptation.

Adding few pronunciation variants through phonological rules (model M2.A5) provides better results for almost all language groups, hence proving their importance in such a recognition task, even when models are adapted using non-native speech data.

As for the results obtained with models relying on units initially trained on foreign data, their behavior after task and all accents adaptation are somewhat similar to the behavior of the results obtained with the standard generic models: the introduction of units corresponding to the adaptation of French phonemes on foreign data (model M4.A5) leads to better recognition performance than the introduction of standard units of foreign languages (model M3.A5). Moreover, using the foreign adapted French units and restricting the variants to those that are the most frequently observed on forced alignments of the adaptation set data, remains the best performing solution, even after task and all accents adaptation (model M5.A5).

However, the recognition performance obtained for each speaker language group with model M5 (selection of units) compared to performance obtained with model M4 (all variants) does not vary in exactly the same way for adapted models (Figure 3) and generic models (Figure 6). This is due to the adaptation on accented data, which adjust the acoustic parameters of the modeling without using any a priori knowledge about the type of accent they are supposed to model. Moreover, a mixed of non-native accents (adaptation set) are used in the adaptation process. Nevertheless overall results are similar, the selection of units leads to a small performance degradation on average on the group of Spanish, English and German speakers (group of languages corresponding to the added units), while performance improvements are observed for French speakers as well as speakers from other languages. Global results are summarized in Table 3.

Table 3 – Summary of results (error rates) with adapted acoustic models.

	French	Span., Engl. & Germ.	Other language groups
	1227 utt.	3106 utt.	4322 utt.
M1.A5: Baseline, adapted on all types of accents	3.2 % (+/- 1.0 %)	5.6 % (+/- 0.8 %)	5.9 % (+/- 0.7 %)
M4.A5: phon. rules & all units; adapted on all types of accents	4.6 %	5.0 %	5.6 %
M5.A5: phon. rules & select. units; adapted on all types of accents	**4.2 %**	**5.2 %**	**5.2 %**

5.2 Subset of Accents

This second set of experiments aims at studying the impact of the task and accent adaptation when only subsets of accents are present in the adaptation data. In order to do so, the adaptation set was split into several subsets, according to the same language criteria as defined in section 2.2 for the test set: data uttered by French speakers only; data from Spanish, English and German speakers ("EsEnDe" subset); and data from speakers of other language groups.

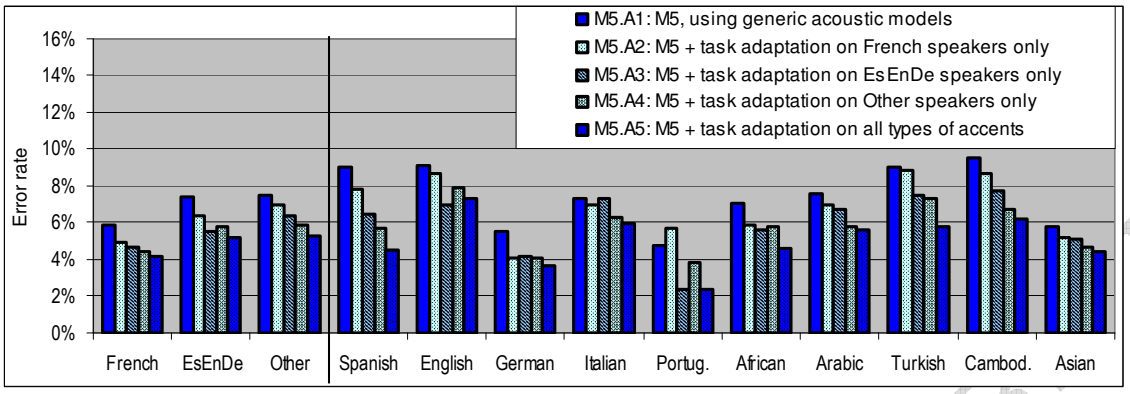


Figure 7 – Error rates for each language group for the best modeling variant, after various accents adaptation.

Results obtained for the best modeling variants (i.e. model M5: application of phonological rules and selection of added foreign-adapted French units) are reported in Figure 7. Results show that the adaptation on French speakers provides some improvement for almost all the language groups. The adaptation on data from non-native speakers (for example models M5.A3 and M5.A4) also provides some improvement for the French group data. This improvement can be considered as the effect of the task adaptation, as recording conditions were similar for French and for non-native speakers.

The adaptation using Spanish, English and German speakers provides noticeable improvements for these speaker language groups. The very interesting point is that such a limited accent adaptation yielded also significant recognition performance improvements on the other speaker language groups. Nevertheless, the adaptation on the subset "Other" reveals that the best results are still obtained when the adaptation and testing sets correspond to the same type of accents. Finally, using the full adaptation set, that is all available types of accents, leads to the best recognition performance. Global results are summarized in Table 4.

Let us also recall that the modeling relies on multiple sets of units for each phoneme, and that the percentage of usage of each of them (along the best Viterbi alignment path) depends on the speaker language group. However, in every case the amount of usage of native French units is important (Bartkova and Jouvet, 2004b). Hence native French units were adapted in each case, which may explain part of the systematic performance improvement observed after adaptation whatever subset was used for adaptation. Also as the frequency of usage vary according to the speaker language group, the more frequently used models on a given group were better adapted with speech data from that group (than with speech data of another group), which resulted in better performance improvements for speakers of the corresponding group.

Table 4 – Summary of results (error rates) for various non-native accent adaptations.

	French	Span., Engl. & Germ.	Other language groups
	1227 utt.	3106 utt.	4322 utt.
M5.A1: phon. rules & select. units; generic models	5.9 % (+/- 1.3 %)	7.4 % (+/- 0.9 %)	7.5 % (+/- 0.8 %)
M5.A3: phon. rules, select. units; adapted on EsEnDe accents	4.7 %	5.6 %	6.4 %
M5.A5: phon. rules & select. units; adapted on all types of accents	**4.2 %**	**5.2 %**	**5.2 %**

6 CONCLUSION

In this paper, recognition of non-native speech, comprising multiple foreign accents, was analyzed. Modeling pronunciation variants proved to be necessary for handling non-native speech variabilities. The application of phonological rules helped handling the replacement of some phonemes by others. Furthermore, introducing models of phonemes in foreign languages provided 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. A new approach, based on the adaptation of models of French phonemes on foreign speech data, was introduced, and it proved to be effective and 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, for example, that foreign models were more frequently used for vowels than for consonants. Therefore automatic selection methods were developed. One of them consisted in keeping for each phoneme only the variants, among the adapted units, that were the most frequently used on a development set. This approach reduced the amount of variants that were introduced, and led to the best and most homogenous recognition results across various speaker language groups.

The adaptation of the acoustic models on various sets of non-native speech was also studied. Results showed that detailed models, including phonological rules and extra units, benefited more from the task and accent adaptation process than baseline standard models used for native speech recognition. Moreover, after adaptation on a limited set of foreign accents, recognition performance was improved even on other types of foreign accents. This interesting result proves that the detailed modeling, relying on the use of phonological rules and models of target phonemes adapted on a few foreign languages, is useful, and takes benefit of the available task accented data through classical adaptation processes, even if all accents are not present in the adaptation data.

ACKNOWLEDGEMENTS

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.

REFERENCES

Aalburg, S., Hoege. H., 2004. Foreign-accented speaker-independent speech recognition. In: Proc. ICSLP'04, Int. Conf. on Spoken Language Processing, Jeju Island, Korea, October 4-8, pp. 1465-1468.

Adda-Decker, M., Lamel, L., 1998. Pronunciation variants across systems, languages and speaking styles. In: Proc. ISCA Workshop on Modelling Pronunciation Variations for Automatic Speech Recognition, Rolduc, the Netherlands, May 4-6, pp. 1-6.

Arslan, L. M., Hansen, J. H. L., 1996. Language accent classification in American English. Speech Communication 18, pp. 353-367.

Bartkova, K.,Jouvet, D., 1999. Language based phone model combination for ASR adaptation to foreign accent. In: Proc. ICPhS'99, Int. Conf. on Phonetic Sciences, San Francisco, CA, USA, August 1-7, pp. 1725-1728.

Bartkova, K., Jouvet, D., 2004a. Multiple models for improved speech recognition for non-native speakers. In. Proc. SPECOM'04, Int. Conf. on Speech and Computer, Saint Petersburg, Russia, September 20-22, pp. 22-28.

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

Bartkova, K.,Jouvet, D., 2006. Using multilingual units for improved modeling of pronunciation variants. In: Proc. ICASSP'2006, IEEE Conf. on Acoustics, Speech and Signal Processing, Toulouse, France, May 15-19.

Beattie, V., Edmondson, S., Miller, D., Patel, Y., Talvola, G., 1995. An integrated multidialect speech recognition system with optional speaker adaptation. In: Proc. Eurospeech'95, Eur. Conf. on Speech Communication and Technology, Madrid, Spain, September 18-21, pp. 1123-1126.

Bonaventura, P., Gallochio, F., Mari, J., Micca, G., 1998. Speech recognition methods for non-native pronunciation variants. In: Proc. ISCA Workshop on Modelling Pronunciation Variations for Automatic Speech Recognition, Rolduc, the Netherlands, May 4-6, pp. 17-22.

Cincarek, T., Gruhn, R., Nakamura, S., 2004. Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models. In: Proc. ICSLP'04, Int. Conf. on Spoken Language Processing, Jeju Island, Korea, October 4-8, pp. 1509-1512.

Delattre, P., Sur les origines celtiques de la prononciation française. In: Studies in French and Comparative Phonetics, pp. 215-217.

Draxler, C., Burger, S., 1997. Identification of regional variants of high German from digit sequences in German telephone speech. In: Proc. Eurospeech'97, Eur. Conf. on Speech Communication and Technology, Rhodes, Greece, September 22-25, pp. 747-750.

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

Fung, P., Liu, W.K., 1999. Fast accent identification and accented speech recognition. In: Proc. ICASSP'99, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Phoenix, Arizona, USA, March 15-19, pp. 221-224.

Goronzy, S., Rapp, S., Kompe, R., 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Communication 42, pp. 109-123.

He; X., Zhao; Y., 2003. Fast model selection based speaker adaptation for nonnative speech. IEEE Trans. on Speech and Audio Processing 11, pp. 298–307.

Huang, C., Chen, T., Li, S., Chang, E., Zhou, J., 2001. Analysis of speaker variability. In: Proc. Eurospeech'01, Eur. Conf. on Speech Communication and Technology, Aalborg, Denmark, September 3-7, pp. 1377–1380.

Humphries, J. J.; Woodland, P. C.; Pearce, D., 1996. Using accent-specific pronunciation modelling for robust speech recognition. In: Proc. ICSLP'96, Int. Conf. on Spoken Language Processing, October 3-6, pp. 2324-2327.

Jouvet, D., Bartkova, K., Monné, J., 1991. On the Modelization of Allophones in an HMM based Speech Recognition System, In: Proc. Eurospeech'91, Eur. Conf. on Speech Communication and Technology, Genoa, Italy, September 24-26, pp. 923-926.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Yu, X., Zhang, S., 2001. What kind of pronunciation variation is hard for triphones to model?. In: Proc. ICASSP'01, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah, USA, May 7-11, pp. 577-580.

Kubala, F.; Anastasakos, A.; Makhoul, J.; Nguyen, L.; Schwartz, R.; Zavaliagkos, E., 1994. Comparative experiments on large vocabulary speech recognition. In: Proc. ICASSP'94, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Adelaïde, Australia, April 19-22, pp. 561-564.

Kumpf, K., King, R.W. 1996. Automatic accent classification of foreign accented Australian English speech. In: Proc. ICSLP'96, 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, pp. 1740-1743.

Lawson, A. D., Harris, D. M., Grieco, J. J., 2003. Effect of foreign accent on speech recognition in the NATO N-4 Corpus. In: Proc. Eurospeech'03, Eur. Conf. on Speech Communication and Technology, Geneva, Switzerland, September 1-4, pp. 1505-1508.

Lin; X., Simske, S., 2004. Phoneme-less hierarchical accent classification. In: Conf. Record of the 38th Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA, USA, November 7-10, pp. 1801-1804.

Liu, W. K., Fung, P, 2000. MLLR-based accent model adaptation without accented data. In: Proc. ICSLP'00, Int. Conf. on Spoken Language Processing, Beijing, China, October 16-20, pp. 738-741.

Livescu, K.; Glass, J.; 2000. Lexical modeling of non-native speech for automatic speech recognition. In: Proc. ICASSP'00. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 5-9, pp. 1683-1686.

Mokbel, C., Colling, O., 1999. Incremental enrollment of speech recognizers. In: Proc. ICASSP'99. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Phoenix, USA, March 15-19, pp. 453-456.

Mokbel, C., Jouvet, D. & Monné, J. Deconvolution of telephone line effects for speech recognition. Speech Communication 19, pp. 185-196

Raux, A., 2004. Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition. In: Proc. ICSLP'04, Int. Conf. on Spoken Language Processing, Jeju Island, Korea, October 4-8, pp. 613-616.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication 29, pp. 225-246.

Teixeira, C., Trancoso, I., Serralheiro, A., 1996. Accent identification. In: Proc. ICSLP'96, Int. Conf. on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, pp. 1784-1787.

Uebler, U., Boros, M., 1999. Recognition of non-native german speech with multilingual recognizers. In: Proc. Eurospeech'99, Eur. Conf. on Speech Communication and Technology, Budapest, Hungary, September 5-9, pp. 911-914.

Valée, N., Boë, L. J., Schwartz, J. L., 1990. Système vocalique: topologie et tendances universelles. In: Proc. JEP'90, Journées d'Etudes sur la Parole, Montréal, Québec, Canada, May 28-31, pp.32-36.

Van Compernolle, D., Smolders, J., Jaspers, P., Hellemans, T., 1991. Speaker clustering for dialectic robustness in speaker independent speech recognition. In: Proc. Eurospeech'91, Eur. Conf. on Speech Communication and Technology, Genoa, Italy, September 24-26, pp. 723-726.

Van Compernolle, D., 2001. Recognizing speech of goats, wolves, sheep and … non-natives. Speech Communication 35, pp. 71-79.

Witt, S., Young, S., 1999. Off-line acoustic modelling of non-native accents. In: Proc. Eurospeech'99, Eur. Conf. on Speech Communication and Technology, Budapest, Hungary, September 5-9, pp. 1367-1370