ACOUSTIC-PHONETIC MODELING OF NON-NATIVE SPEECH FOR LANGUAGE IDENTIFICATION

R. Wanneroy[1]*[*] *E. Bilinski*[2], *C. Barras*[1], *M. Adda-Decker*[2], *E. Geoffrois*[1].

[1]DGA/CTA/GIP, 16 bis av. Prieur de la Côte d'Or, F-94114 Arcueil cedex
[2] LIMSI-CNRS, bat. 508, BP 133, F-91403 Orsay cedex

Abstract

The aim of this paper is to investigate to what extent non native speech may deteriorate language identification (LID) performances and to improve them using acoustic adaptation. Our reference LID system is based on a phonotactic approach. The system makes use of language-independent acoustic models and language-specific phone-based bigram language models. Experiments are conducted on the SQALE test database, which contains recordings from English, French and German native speakers, and on the MIST database, which contains non-native speech in the same languages uttered by Dutch speakers. Using 5 seconds of telephone quality speech, language identification error rate amounts to 19% for native speech and to 31% for non-native speech, thus yielding about 60% relative error rate increase. Eventually we propose to improve non-native language identification by an adaptation of the acoustic models to the non-native speech.

1 INTRODUCTION

In the field of automatic speech processing, intensive research activities have been devoted to speech recognition and transcription. With the growing interest in multilinguality and multilingual systems, language identification (LID) has become a research area of its own [3, 6]. In a multilingual context however speakers may use foreign languages for communication. Under such conditions, i.e. dealing with non-native speech input, system performances are known to decrease. Yet systematic evaluations of such degradation and research efforts to minimize them are still to be fostered.

Various information sources can be exploited in order to identify a given language: acoustic, phonemic, phonotactic, lexical, etc. For each information level specific resources and corpora are required for the languages to be modeled. In most LID approaches only acoustic-phonetic and phonotactic models are used. Given the spectral distorsions commonly observed in non-native speech, performance will degrade when using acoustic-phonetic and phonotactic models trained from native speech.

Studying the impact of non-native speech on LID requires appropriate test material. Ideally a multilingual native speaker database and a multilingual non-native speaker database are required. Both corpora should be similar in style and recorded in comparable acoustic conditions. To our knowledge the MIST database is the first multi-lingual corpus gathering non-native speech; it contains recordings in English, French and German from Dutch speakers. Similar native speech material is provided by the multilingual corpora produced within the LE-SQALE project [4].

In the following, we describe the LID system used for the experiments. We present baseline LID results on native speech using the SQALE test database, and results on non-native speech using the MIST database; by means of these experiments we measure the impact of native versus non-native speech on LID error rates. Finally we investigate the effectiveness of acoustic model adaptation to handle non-native speech.

2 LID SYSTEM

The LID system used in the experiments is based on a phonotactic approach, with a single language-independent acoustic-phonetic decoder. This approach was chosen because, compared to language-specific acoustic modeling, it allows easier extension of the system to new languages, as there is no need

[*] financed by CS (Communications & Systems) company.

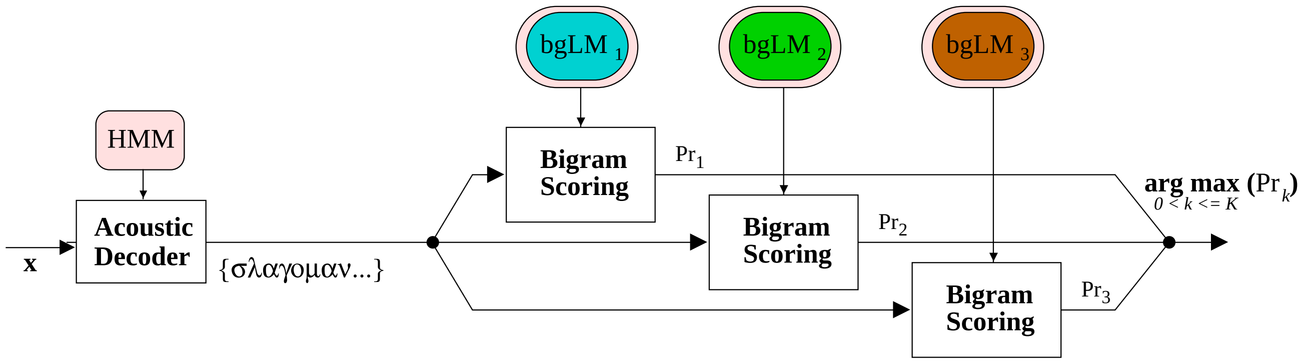

Figure 1: LID system using language-independent acoustic models and phone-based bigram language models

of specific phonetic knowledge of the new language or of a phonetic labelling of the training databases. The phonotactic approach generally requires longer test segments to obtain optimal results as compared to acoustic-phonetic approaches. Previous work showed that the phonotactic approach LID results significantly improve when the test segment length goes from 10s to 45s [5].

The system is more extensively described in another article [1], where it is referenced as LI_HC (language-independent hierarchically clustered phone set). It is illustrated in Figure 1. It uses one single language-independent phone recognizer to label the speech input. The phone sequence output by this phone recognizer is then scored with language-dependent phonotactic models approximated by phone bigrams. The language providing the highest probability is hypothesized.

2.1 Training database

The LID system was trained using the IDEAL corpus, which is a multi-language telephone speech corpus designed to support research on LID [2]. This corpus contains a large amount of speech (about 19 hours per language). The different languages were collected under the same conditions, and native speakers were recruited in their home countries. Data have been recorded for British English, Spanish, French and German. All speakers called the LIMSI data collection system ensuring the same recording conditions for the entire corpus. The IDEAL corpus contains about 300 calls for each language (i.e., international calls from native U.K., Spanish, and German speakers and national calls from native French speakers), 250 of them being used for acoustic and phonotactic model estimation.

The calling script was designed to cover a variety of data types: 12 questions to elicit precise responses (7 general questions concerning the call and caller, and 5 prompts asking for times, dates, days of the week and months of the year), 18 items containing predefined texts to read, and 6 questions aimed at collecting spontaneous speech. The acoustic models were trained on all types of material, and the phonotactic models on the spontaneous speech part.

2.2 Front-end processing

The front-end processing consists in 12 MFCC plus the energy, augmented by their first and second order derivatives, i.e. a total of 39 coefficients every 10 ms. The same setting was used for processing test data, except that signal frequencies over 3.5 kHz were cut in order to be consistent with the training database which contains only narrow-band telephone speech.

2.3 Acoustic models

250 calls from IDEAL (about 9000 sentences, containing up to 13 hours of speech for each language) have been used for acoustic model training. First, 4 language-specific phone sets for English, French, German, and Spanish were trained. All acoustic models are three-state continuous density HMM of context-independent phones. Then a single multi-lingual set of 91 monophone models was obtained by an agglomerative hierarchical clustering of these 4 phone sets, using a measure of similarity between phones [1]. This phone set has proven to allow effective extension to new languages [5].

2.4 Phonotactic models

Phonotactic models were estimated on the spontaneous speech part of the 250 training calls which accounts for about 15% of the IDEAL corpus. For each language, an acoustic-phonetic decoding of the training database was performed using the multilingual phone set. The decoded phone strings are then used to estimate language-dependent bigram models for English, French and German.

3 TEST CORPORA

Experiments were conducted on the SQALE and MIST databases for LID results on native and non-native speech, respectively.

3.1 SQALE database

The development and test data of the SQALE project [4] were used for the native speech experiments. The 4-language (French, British and American English, German) speech database contains 400 sentences per language from 40 speakers, plus some diagnostic sentences which were not used in our experiments. Within the SQALE project the test sentences were chosen to give a reasonable spread of difficulty as determined by sentence length and perplexity. French, English (British or American) and German speakers were recorded reading newspaper texts from Le Monde, Wall Street Journal and Frankfurter Rundschau, respectively.

3.2 MIST database

The MIST database was developed by the TNO Human Factors Research Institute to support research in multi-linguality and non-native speech. 74 native Dutch speakers (52 male, 22 female) uttered 10 sentences in Dutch, and also for most of them in English, French and German: 5 sentences per language identical for all speakers and 5 unique sentences per language and per speaker. The text sources are the same as for the SQALE project concerning English, French and German plus the Dutch NRC/Handelsblad. We only used unique sentences for evaluation of non native speech because identical sentences are not phonetically balanced over time. Finally, the selected part of the MIST database contains about 300 sentences per language.

4 EXPERIMENTAL RESULTS

We present LID error rates for each language as a function of sentence duration. Every second, the system takes a decision on the speech segment decoded so far. In order to reduce duration variability due to pauses and hesitations, the silences labelled by the recognizer are discounted from the sentence duration. For both test corpora, mean sentence duration is about 6 seconds. Few sentences are more than 8s long, and no significant LID results were obtained for segment durations over this duration.

4.1 Results on native speech

For the four SQALE languages Figure 2 provides identification results on a second per second basis. Concerning 5 second segments, the global error rate amounts to 19%. This global rate does not show the disparity between languages; indeed, error rates of 3%, 20%, 24% and 28% are achieved for French, German, American and British English speech respectively. Although the English phonotactic model has been trained over British English data, the error rate is higher than for American English. For all durations, results on French are significantly better than on the other languages. This might be attributed to the difference between French national and international telephone networks.

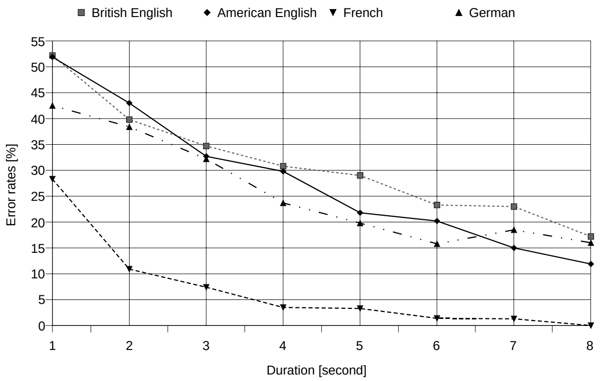

Figure 2: LID error rates for the four native language task (SQALE database) as a function of segment duration.

4.2 Results on non-native speech

Similar experiments were conducted on non-native speech. The identification results using the three non-native MIST languages are illustrated in Figure 3. Results on German appear to be significantly worse than on the two other languages. Especially when displaying the error rate against segment duration, a much slighter slope could be observed for German,

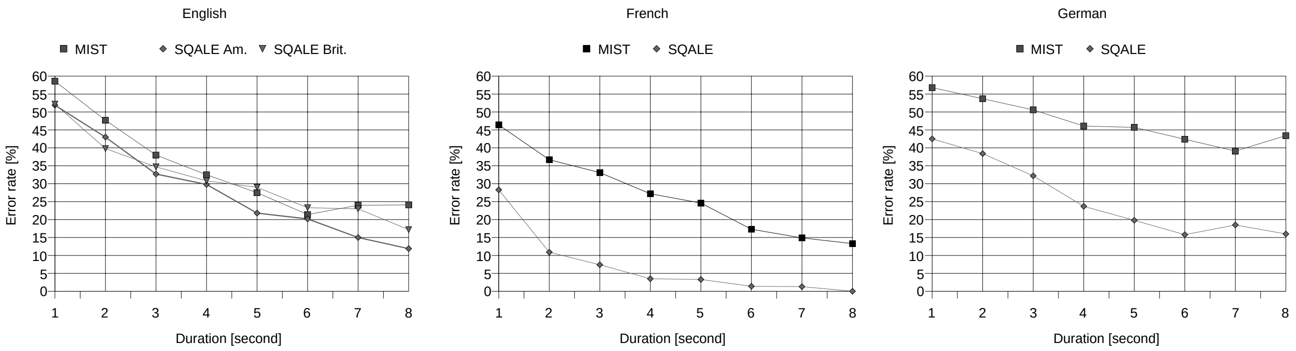

Figure 4: LID error rate comparison between native and non-native speech for English (American or British), French and German as a function of segment duration.

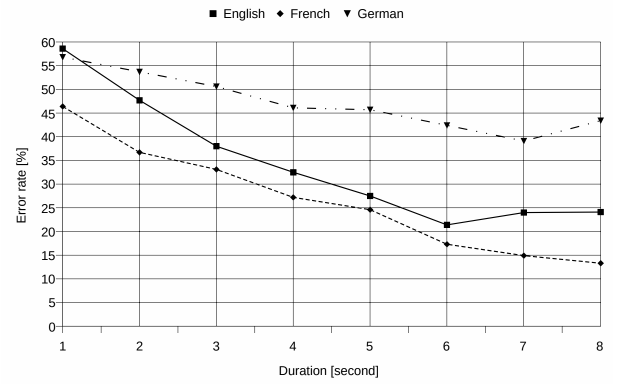

Figure 3: LID error rates for the three non-native language task (MIST database) as a function of segment duration.

as compared to French or English. On 5 second segments, LID error rates for non-native French, English and German are 23%, 27% and 44%, respectively. The global LID error rate of the three non-native languages is 31%.

4.3 Comparison between native and non-native speech

The comparison of the identification results for native and non-native speech for each language is illustrated in Figure 4. For French and German, the non-native Dutch accent increases the error rates as expected. But American, British and non-native English obtain roughly the same error rates for each segment duration. The English phonotactic model seems to be more robust with respect to accent variation. Another more linguistically motivated conclusion consists in suggesting that Dutch speakers are best in speaking English as compared to French and German. For 5 second segments, the global error rate amounts to 19% for native speech and to 31% for non-native speech, showing a 60% relative error rate increase (cf. Table 1).

Table 1: Per language and global LID error rates on native speech (SQALE database) and non-native speech (MIST database) for 5 seconds of speech.

	SQALE	MIST	relative increase
British English	28%	27%	×1
American English	24%		×1.1
French	3%	23%	×8
German	20%	44%	×2.2
Global rate	**19%**	**31%**	**×1.6**

4.4 Adaptation of acoustic models

Better results on non-native speech should be obtained after an adaptation of the LID system to the new conditions. Given the size of the available non native speech material (the MIST test database), an adaptation of the phonotactic models does not seem possible, and only acoustic models adaptation was tested. For a better use of the available data, the non-native MIST data were jack-knifed to produce 5 sets of adapted phone models. Each non-native sentence of the adaptation subset is aligned with the original prompt using the language-dependent acoustic models and produces a phone segmentation which is converted into the language-independent phone set. The acoustic models (including means, variances and weights of gaussians) are adapted towards the non-native acoustic realization of the phones. A weight-

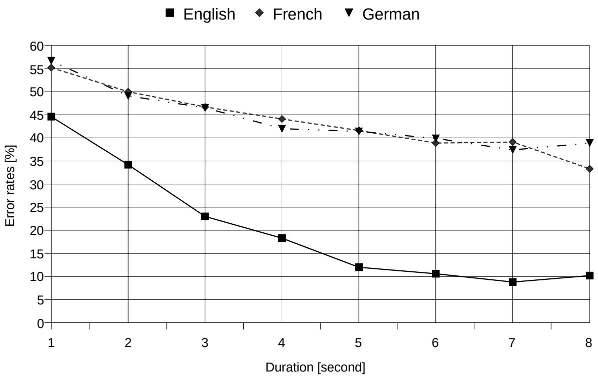

Figure 5: LID error rates for the three non-native language task (MIST database) as a function of segment duration after adaptation of acoustic models.

ing factor allows to control the degree of adaptation. The LID system with adapted acoustic models is finally tested on the left-out fifth of the database.

Figure 5 shows the LID error rates after adaptation of the acoustic models. On 5 second segment, LID error rates of 41%, 14% and 41% are achieved for non-native French, English and German respectively. This leads to an important improvement for English (14% with adaptation vs. 27% without adaptation, an even better result than for native speech), a slight improvement for German (41% with adaptation vs. 44% without adaptation) and a serious degradation of results for French (41% vs. 23%). However the global LID error rate for the three non-native languages does not change significantly after the adaptation.

5 CONCLUSIONS

Experiments have been carried out with a phonotactic-based approach LID system on a 3-language task using native and non-native speech (SQALE, MIST corpora).

Using 5 seconds of telephone quality speech, LID error rate increased from 19% for native speech to 31% for non-native speech. Given the limited amount of test data, the test segment duration has been limited to a maximum length of 8 seconds, which stays far away from the typical durations (30s and more) for which the phonotactic LID approach performs best.

Adaptation of the acoustic model sets did not allow to significantly reduce the error rate on non-native speech. However we may notice that acoustic modeling is most effective for English for which only a small degradation of the results is observed on non-native speech. Using the phonotactic approach, adaptation of the phonotactic models should be more efficient, but it could not be tested with the databases involved.

Needs for further investigation are clear. Studying the effects of non-native speech on LID requires bigger databases with more utterances of a longer duration, more languages and various foreign accents. The development cost of such resource is of course the main question. But MIST database, even if only devoted to Dutch accent over a few European languages, was clearly an excellent starting point for the study of non-native speech, especially because of its matching with the already studied SQALE database.

Acknowledgements

The authors wish to thank TNO for sharing the MIST database.

References

[1] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel, "Language Identification with Language-Independent Acoustic Models", *Eurospeech'97*, pp. 55-58, Rhodes, Sept. 1997.

[2] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J. Gangolf, J.L. Gauvain, "A Multilingual Corpus for Language Identification", *1st Int. Conf. on Language Resources and Evaluation*, pp. 1115-1122, Granada, May 1998.

[3] Y.K. Muthusamy, E. Barnard, R.A. Cole, "Reviewing Automatic Language Identification", *IEEE Signal Processing Magazine*, Oct. 1994.

[4] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual large vocabulary speech recognition: the European SQALE project", *Computer Speech and Language*, **11**, pp. 73-89, 1997.

[5] D. Matrouf, M. Adda-Decker, J.L. Gauvain, L. Lamel, "Comparing different model configurations for language identification using a phonotactic approach", *Eurospeech'99*, Budapest, Sept. 1999.

[6] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. on SAP*, **4**(1), pp. 31-34, Jan. 1996.