ACCENT IDENTIFICATION

Carlos Teixeira, Isabel Trancoso and António Serralheiro

INESC/IST
Instituto de Engenharia de Sistemas e Computadores
INESC sala 236, Rua Alves Redol 9,
1000 LISBOA Apartado 13069, PORTUGAL
Phone: +351.1.3100314
Fax: +351.1.3145843
Email: cjct@inesc.pt

ABSTRACT

Foreign accent identification is a new challenging problem closely related to other relatively recent fields of the multilinguality area such as dialect identification and language identification. This paper describes an automatic identification system for English accents from 6 different European countries. The approach is based on a parallel set of ergodic nets with context independent HMM units. The ergodic topology was also substituted by pronunciation transcription constraints in order to integrate accent specific automatic word recognisers. Considering the complexity of the task, the results can be considered encouraging for further research.

1. INTRODUCTION

Nowadays, Automatic Speech Recognition is moving from the laboratory to the market mainly through applications running on the telephone network and on personal computers. Most of these applications have been developed for one language and their use by non-native speakers causes a significant drop of performance compromising the usability of such systems[10].

Non-native accents can be also found in travelling and tourist centers or automatic international phone call services. Since English is probably the mostly used second language in the world, many speakers use it when addressing such services abroad, expecting some immediately feedback. If the listener is able to identify the accent, he may be able to find a suitable attendant knowing the first language of the specific client. For an automatic system, this means selecting an automatic speech recogniser for that first language or, if there is no such recogniser, another one adapted for the given English accent.

Non-native accent identification is a new challenging problem closely related to other new fields from the multilinguality area. One of the most important ones is Language Identification (LI). In this field, there is a vast amount of knowledge for each specific language (phoneme inventories, for instance) which allows us to discriminate between languages. A related field is automatic Dialect Identification (DI), a relatively recent area of research which is very close to non-native accent identification. Dialectal differences are often proudly marked and native speakers do not generally attempt to conform to a standard variant. Non-native speakers, on the other hand, show different degrees of reading competence and pronunciation competence [7], that is, their knowledge of the grapheme-to-phoneme conventions of the foreign language may vary a lot, as well as their ability to pronounce sounds which are not part of their native sound inventory.

The growing interest in the area of non-native accents is demonstrated by the recently collected TED corpus (Transnational English Database), containing recordings from EUROSPEECH'93 presentations, and also, for instance, by the work reported in [2], and [3]. The first authors present results with continuous speech samples of English by speakers whose first language is Arabic, Chinese and Australian. The second authors, on the other hand, present results with isolated words spoken in English by speakers with English, Chinese, Turkish and German accents.

Our first restrict recognition experiments with non-native accents indicated a drop of aprox. 15% for non-native English speakers when using a recogniser not trained with their specific accent [10]. At the same time, Brousseau & Fox [1] indicated similar results for different dialects of English as well as French. Our first approach consisted of training whole word models with material spoken by a representative population of non-native speakers, as well as native ones (of English, in our case). The results obtained were satisfactory. Later experiments [11] used two sets of models, one trained with native speakers and another with non-native ones. The performance improved slightly, though at the cost of doubling the computation time. This experiment also provided figures of the discrimination capability of such a system, in terms of native and non-native speakers distinction, and indicated a stronger separation of speakers according to their gender than according to their accent.

The present work is based on a relatively small corpus of about 200 isolated words, with speakers from 6 different countries. This corpus attempts to illustrate the main problems caused by the use of an automatic speech recogniser by non-native speakers, in typical small vocabulary applications and is described in Section 2. Section 3 describes the baseline recogniser as well as the procedures for obtaining accent-specific and globally trained phone models. Section 4 introduces the accent identification system, its integration with a speech recognition system, and the corresponding results. Conclusions and directions for future research are discussed in the final section.

2. MULTI-ACCENT EUROPEAN ENGLISH CORPUS

A spoken corpus was collected for five small vocabulary applications in English[10], in the scope of the SUNSTAR European project. Two of this applications were developed for the telephone network and the others for office environment. The Portuguese accent (*pt*) was recently added to the original set representing 5 accents: Danish (*da*), German (*de*), British (*en*), Spanish (*es*), and Italian (*it*). There are 20 speakers (approx. 10 male and 10 female) of each accent, each one repeating two times a vocabulary of 200 English isolated words.

The corpus was recorded and orthographically labelled according to the standards recommended in the SAM project, and later downsampled and filtered to telephone line bandwidth. All the material was heard at least once in order to drop defective utterances. The full corpus was data compressed to fit into a CDROM.

Most of the speakers in the 5 non-English speaking countries were closely involved with the research labs which performed the data collection. Hence, they may not be considered a representative sample of their country population in terms of knowledge of English, since they are probably more experienced with reading, listening and even talking English than most of their compatriots. However, the speakers seldom lived abroad in an English speaking country. We therefore considered them representative of potential users for systems incorporating automatic accent identification. Beside these general comments, there is little information about each speaker, other than his/her sex, first language and age.

Very informal listening tests have been performed using a subset of the same corpus, consisting of isolated words spoken by 6 female speakers, one from each country. 5 Portuguese listeners and 1 French one have heard sequences of words pronounced by the same speaker, until they could make a guess about her accent. All the listeners have frequently participated in international projects and conferences, and were therefore accustomed to hearing English spoken by foreigners. Two main conclusions were derived from this small test: (1) it was very difficult for the listeners to make a guess without hearing at least some 6 to 12 words from each speaker; and (2) the British English accent was always identified correctly. We have also observed other facts which may be very dependent on the selected speakers and which may, therefore, be not generalizable: (3) the Italian and Spanish accents were often confused with each other; (4) the German and Danish accents were also confused either with each other or with English, though not so often; (5) the Portuguese accent was harder to characterize. This informal listening test will help us design a proper human benchmark test which may yield interesting conclusions about the difficulty of the accent identification task.

It was also interesting to notice that, despite the fact that, for each country, most speakers were recruited within the same personnel department, the reading and pronunciation competence shown by the non-native speakers varied very much, even within a single accent. Reading errors varied usually between 5 and 10% and were noticed mainly with vowels (i.e. the vowel *i* is sometimes erroneously pronounced as a diphthong when it should be pronounced as [i] or vice-versa). Pronunciation errors are mainly due to the difference in phoneme inventories between the native language and English, in our case. As an example, the sound [θ], which does not occur in Portuguese, is often approximated by a native one - [s].

3. THE BASELINE RECOGNISER

For the purpose of training models and following the results of the previous work [11], the spoken corpus described above was divided into two sets, according to gender (male /female). Each of these two sets was further split into training and testing subsets - about 60% for training and the remaining 40% for testing. Finally, in order to obtain vocabulary independent tests, only 25% of the words were used for testing while the remaining were used for training context independent sub-word models.

While context-dependent models are known to provide better performance than context-independent models, the amount of vocabulary available for training is too small in terms of diphone or triphone coverage to cover a test vocabulary with a significant size (ex: bigger than 10 words).

Since the corpus contained mostly isolated words and a few compound words (i.e. *alarm-call*), we decided not to use prosodic features in this study. However, there are very pronounced rhythmic differences among accents, which we are planning to explore in future work.

The signal pre-processor introduces pre-emphasis and a Hamming window of 30ms is shifted every 10ms in order to perform an LPC analysis of order 12 and compute 8 liftered cepstra and the corresponding delta cepstra coefficients.

The pronunciation lexicon distributed with the TIMIT database was adopted, with slight modifications. We have used 48 context independent phone models, with implicit insertion/deletion modelling in stop-closure pairs [5], and 3-state CHMM linear topology. Initial models are built using a linear segmentation procedure, followed by a Viterbi alignment and later reestimated using the Baum-Welsh algorithm. Six multiple mixtures are obtained by iterative splitting, in a process very similar to the one reported in [9], with the HTK package, on which most of the system was developed. Recognition was done using Viterbi decoding.

The first set of recognition results concerned models trained and tested with speakers with the same accent, and is presented in Table 1.

One can notice that best results were obtained with the English accent, which could be expected given the better adequacy of the phonetic transcriptions and smaller dispersion of pronunciations. Worst results were obtained with Italian and Spanish accents.

A second set of tests was done with models trained with all the accents together, for each sex. The recognition scores are presented in Table 2 and were obtained using 6 mixture components, as in Table 1, for the sake of comparison. In fact, slightly better results were obtained with a higher number of mixture components, given the larger

amount of training material.

Accent	Female	Male
da	82.9	75.4
de	85.8	81.8
en	92.0	90.8
es	75.8	74.5
it	68.3	71.8
pt	85.3	85.0

Table 1: Percentage recognition scores using specific recognisers for each gender and accent.

Accent	Female	Male
da	82.5	86.5
de	90.7	93.9
en	92.0	91.7
es	79.3	78.7
it	76.5	73.6
pt	85.0	88.1

Table 2: Percentage recognition scores using one single recogniser per gender for every accent.

This experiment expanded previous results using only 2 accents [10]. As in this previous work, which used whole-word models, the results obtained indicate that a single set of models trained with all the accents can improve the performance obtained with the recognisers trained with the specific accents.

4. AN ACCENT IDENTIFICATION SYSTEM

Accent identification could be used in a three-stage system, in which the first stage would decide about the speaker gender, the second stage would classify the speaker accent, and the final stage would use the recogniser models corresponding to the decisions made in the previous stages.

Concerning the accent identification stage, we used an HMM technique already tested by other researchers in Language Identification tasks [4]. A topology of parallel competing sub-nets was adopted, in which each sub-net consisted of an ergodic net of the full set of phone HMMs trained with the corresponding accent. The sub-net which achieves the higher likelihood is selected. A silence model was trained with non-speech segments of the utterances of most speakers, in order to capture any variations along the different recording environments. This model is used as an alternative branch between compound words as well as in the beginning and end of each word.

The results presented in Tables 3 and 4 were obtained with a system which integrated gender identification with accent identification. This was done simply by adding in parallel both gender accent subnets in the same way as it was done before. This approach does not consider any knowledge about a specific application: it can work with any portion of speech including any vocabulary. Hence, no pronunciation lexicon was used in these experiments.

Gender	Female	Male
Female	96.1	3.9
Male	9.8	90.2

Table 3: Confusion Matrix (%) for gender identification using ergodic sub-nets.

Accent	da	de	en	es	it	pt
da	64.4	2.7	10.7	11.7	2.9	7.6
de	5.5	64.2	6.0	1.5	4.8	18.0
en	4.1	6.2	73.3	3.6	2.9	10.0
es	14.4	7.2	8.1	56.0	1.8	12.5
it	2.9	7.7	2.9	5.9	62.7	17.9
pt	1.9	6.4	12.5	4.7	3.2	71.4

Table 4: Confusion Matrix (%) for accent identification using ergodic sub-nets.

The global score obtained in these tests for gender identification was 92.9%, which shows that this identification stage works reasonably well. For accent identification, the global score was 65.48%. Best results were obtained for English, but off diagonal scores can be as high as 18%.

The previous experiment was performed with no knowledge of the application vocabulary. In the following one, phone transcriptions were integrated. This is equivalent to what can be called a fixed state grammar in contrast with a null grammar used in the previous experiments. With this approach, there is a sub-net for each word, gender and accent, replacing the accent sub-nets in the previous topology. It is now possible to tag each utterance with a word label and one could look at this net as a common phone-based word recogniser with multiple models for each gender and accent. As a by-product it can also do gender and accent identification.

Table 5 presents the confusion matrix for gender identification. The global score is 94.0%. The confusion matrix for accent identification is presented in Table 6. The global score is 65.4%. The addition of the pronunciation lexicon did not yield the expected improvements, which suggests that the single phonetic transcriptions used for each word are not adequate for this task.

Gender	Female	Male
Female	95.7	4.3
Male	7.5	92.5

Table 5: Confusion Matrix (%) for gender identification using phonetic transcriptions for each word.

Table 7 presents the recognition scores obtained in this experiment. The results are of the same magnitude as the ones presented in Table 1, in which the recogniser had a priori knowledge of the accent and gender of the speaker. This suggests that, although the results of automatic accent identification were not so good, the selection of the appropriate recogniser is effectively done.

Accent	da	de	en	es	it	pt
da	64.8	3.0	9.9	12.0	2.6	7.7
de	7.0	57.1	6.4	2.7	4.8	22.0
en	2.9	6.6	74.1	3.7	2.3	10.4
es	13.6	7.2	5.8	63.0	2.1	8.3
it	4.0	9.9	2.9	8.9	62.1	12.1
pt	2.4	8.9	9.2	4.7	3.1	71.7

Table 6: Confusion Matrix (%) for accent identification using phonetic transcriptions for each word.

Accent	Female	Male
da	83.1	77.1
de	87.7	85.0
en	92.4	89.9
es	77.0	73.2
it	69.8	72.0
pt	85.1	84.0

Table 7: Percentage recognition scores using phonetic transcriptions for each word.

5. CONCLUSIONS AND FUTURE RESEARCH

Automatic identification of non-native accents is a difficult task, as illustrated by the results we have obtained with only 6 different accents. We have tested accent identification in isolation and integrated in a speech recognition system. The latter results have shown that it is preferable to train a recogniser with a mixture of accents.

The current work has used single pronunciation networks per word, a restriction which does not take into account the multiple ways in which a word can be pronounced by a non-native speaker, depending on his reading competence and the differenced between his native phoneme inventory and the English one.

We are currently working on deriving multiple pronunciation networks adapted to given accents and are also considering exploring rythmic cues in order to do a more effective accent identification.

6. ACKNOWLEDGEMENTS

The authors would like to thank Céu Viana (CLUL) for numerous helpful suggestions.

7. REFERENCES

1. J. Brousseau and S. A. Fox. Dialect-dependent speech recognisers for canadian and european french. *Proc. ICSLP, Banff*, 2:1003–1006, 1992.
2. J. Vonwiller C. Blackburn and R. King. Automatic accent classification using artificial neural networks. *Proc. EUROSPEECH, Berlin*, 2:1241–1244, 1993.
3. John Hansen and Levent Arslan. Foreign accent classification using source generator based prosodic features. *Proc. ICASSP, Detroit*, 1:836–839, 1995.
4. L. Lamel and J. L. Gauvain. Language identification using phone-based acoustic likelihoods. *Proc. ICASSP*, 1:293–296, 1994.
5. K. F. Lee. *Automatic Speech Recognition: The Development of the SPHINX SYSTEM*. Kluwer Academic Publishers, Boston, 1989.
6. Deborah Rekart Marc Zissman, Terry Gleason and Beth Losiewicz. Automatic dialect identification of extemporaneous conversational latin american spanish speech. *Proc. ICASSP, Atlanta*, 2:777–780, 1996.
7. Andreas Mengel. Transcribing names - a multiple choice task: mistakes, pitfals and escape routes. *Proc. 1st ONOMASTICA Research Colloquium, London*, pages 5–9, 1993.
8. Isabel Trancoso (on behalf of the Onomastica Consortium). The onomastica interlanguage pronunciation lexicon. *Proc. EUROSPEECH, Madrid*, 1:829–832, 1995.
9. S.J. Young & P.C.Woodland. The use of state tying in continuous speech recognition. *Proc. Eurospeech*, 3:2203–6, 1993.
10. Carlos Teixeira and Isabel Trancoso. Word rejection using multiple sink models. *Proc. ICSLP, Banff*, 2:1443–1446, 1992.
11. Carlos Teixeira and Isabel Trancoso. Continuous and semi-continuous hmm for recognising non-native pronunciations. *Proc. IEEE Workshop ASR*, pages 26,27, 1993.