RECOGNITION OF NON-NATIVE ACCENTS

Carlos Teixeira, Isabel Trancoso and António Serralheiro
INESC/IST
INESC, Rua Alves Redol 9, 1000 LISBOA, PORTUGAL
Phone: +351.1.3100314, Fax: +351.1.3145843, Email: cjct@inesc.pt

ABSTRACT

This paper deals with the problem of non-native accents in speech recognition. Reference tests were performed using whole-word and sub-word models trained either with a native accent or a pool of native and non-native accents. The results seem to indicate that the use of phonetic transcriptions for each specific accent may improve recognition scores with sub-word models. A data-driven process is used to derive transcription lattices. The recognition scores thus obtained were encouraging.

1. INTRODUCTION

Our main concern in this study is to reduce the negative effects caused by non-native accents in automatic speech recognition. These effects account for a significant drop of word recognition scores when using a recogniser trained with material spoken with a native accent. Collecting large enough corpora for each non-native accent is generally not feasible. However, this problem turns out to be more complex, since even a recogniser trained with speech material from a specific non-native accent, still achieves relatively low recognition scores for speakers with that same accent, given the larger range of pronunciations among non-native speakers.

Our first recognition experiments with non-native speakers of English and whole-word models indicated a drop of approximately 15% when using a recogniser trained for native speakers [6]. At the same time, Brousseau & Fox [1] indicated similar results for different dialects of English as well as French.

In our recent research [3], we have tried to evaluate the discriminative capabilities of HMM models in terms of accent identification, an area which has also been the concern of other researchers recently [2]. A topology of parallel competing sub-nets was adopted, in which each sub-net consisted of an ergodic net of the full set of HMM phone models trained with the corresponding accent. The sub-net which achieves the higher likelihood was then selected. This accent identification approach was integrated in a three-stage speech recognition system, in which the first stage decided about the speaker gender, the second stage classified the speaker accent, and the final stage used the recogniser models corresponding to the decisions made in the previous stages. The results, however, have shown that it is still preferable to train a recogniser with a pool of accents.

The spoken corpus used in our research work was collected in the scope of the SUNSTAR European project [6]. The corpus originally comprised five accents of English (Danish (*da*), German (*de*), British (*en*), Spanish (*es*), and Italian (*it*), to which Portuguese (*pt*) was later added. There are 20 speakers (approx. 10 male and 10 female) for each accent, each one repeating two times a vocabulary of 200 English isolated words. The experiments described in this paper used only the male sub-set of this corpus. This sub-set was split into training and testing sub-sets of speakers (60% and 40%, respectively).

Section 2 presents results obtained with whole-word models, which should serve as an upper-bound measure for the remaining vocabulary independent experiments, described in sections 3 and 4. The sub-word models experiments described in section 3 constrain the sub-word models with a single phonetic transcription, derived from a pronunciation dictionary. This restriction 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 differences between his native phoneme inventory and the foreign one. This fact motivated the use of transcription lattices for which a data-driven method will be described in section 4.

2. WHOLE-WORD EXPERIMENTS

In the experiments described in this section, only 25% of the vocabulary available was used for testing. The remaining vocabulary was reserved for the experiments in the following sections, for training vocabulary independent sub-word models. The speech signal was pre-processed to introduce pre-emphasis and a Hamming window of 30ms was shifted every 10ms in order to perform an LPC analysis of order 12. Eight liftered cepstra and the corresponding delta cepstra coefficients were then computed. The word models followed a 10-state CHMM linear topology. Recognition was done using Viterbi decoding.

The first set of recognition experiments used whole-word models trained exclusively with English native speakers. The results are summarised in figure 1a. In this figure (and in the following similar ones) the leftmost points correspond to the recognition scores obtained using the initial models which were built by linear segmentation, followed by a Viterbi alignment. These initial models were afterwards reestimated until convergence was obtained for a maximum of 10 reestimation cycles. The second column of points refers to the models obtained after 4 cycles of embedded reestimation. The two sets of models we have just described had a single Gaussian observation dis-

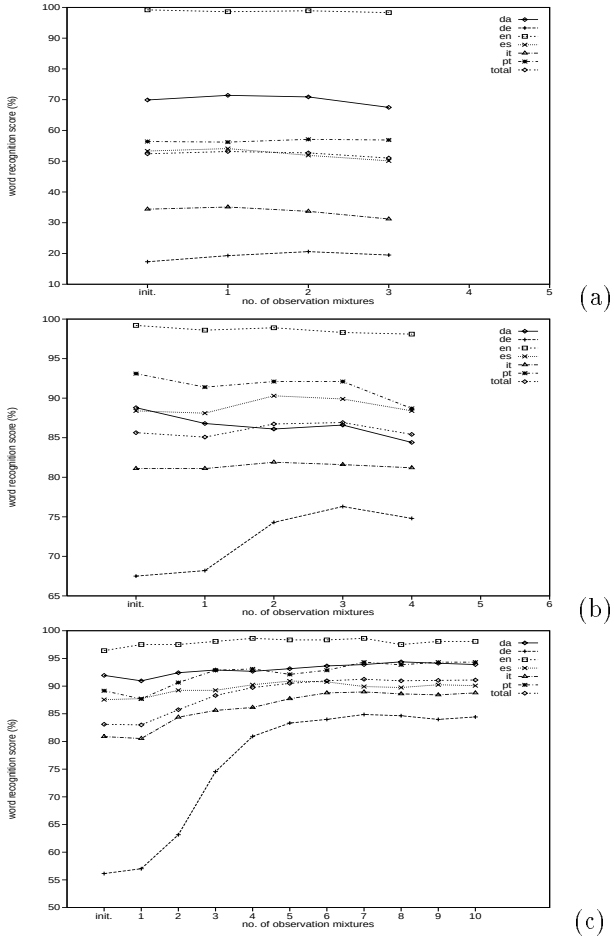

Figure 1. Percentage recognition scores using whole-word models with different number of observation mixtures, trained with: (a) British speakers; (b) speakers from each specific accent; (c) all the speakers from the training corpus sub-set.

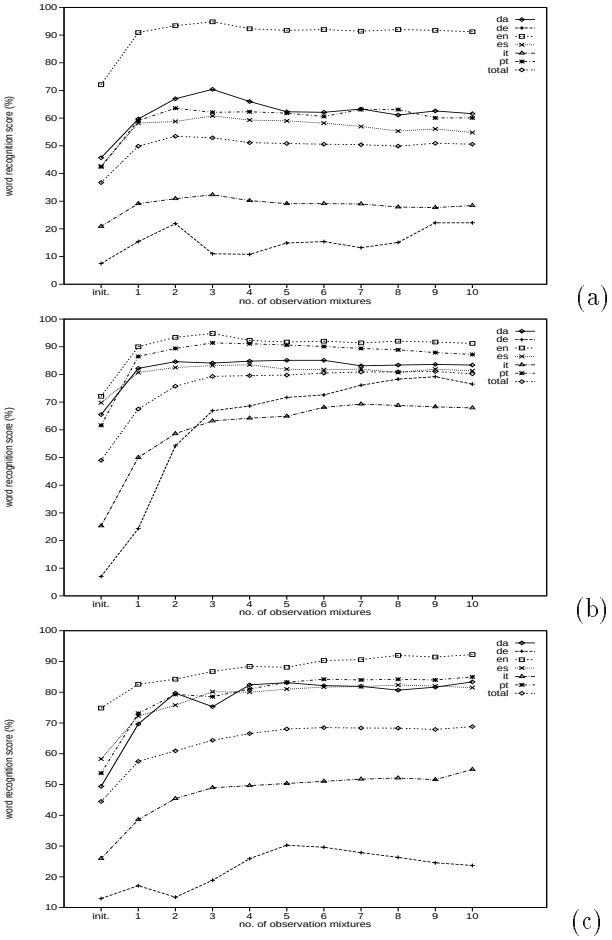

Figure 2. Percentage recognition scores using sub-word models with different number of observation mixtures, trained with: (a) British speakers; (b) speakers from each specific accent; (c) all the speakers from the training corpus sub-set.

tribution. The remaining scores in the figure refer to models with multiple mixtures, which were obtained by iterative splitting, in a process similar to the one reported in [7], with the HTK package, on which most of the system was developed. The line labeled *total* reports the average score obtained overall accents. The best non-native score in figure 1a was obtained by the Danish speakers. Worst results were obtained with Italian and German accents. The vocabulary chosen for testing seems to include particularly hard words for these speakers, since the bad scores were obtained for specific words and not for specific speakers.

The second set of experiments concerned models trained and tested separately with speakers with the same accent (figure 1b). As expected, these specific recognisers provided far better recognition scores when compared with figure 1a (except, obviously, for the British accent).

Non-native speech recognition can be viewed as a speaker independent recognition problem, for which the traditional approach has been pooling all the speech data from as many speakers as possible as if it would belong to a single speaker. The recognition scores obtained using this approach are plotted in figure 1c, which presents the best results for all non-native accents. The British speakers are still holding the best score but the average distance to the non-native speakers is now less than 9%.

3. SUB-WORD EXPERIMENTS

The experiments performed with whole-word models were repeated as a vocabulary independent task using sub-word models (figure 2). Most of the parameters from the whole-word recogniser were used, except the number of states which was set to 3. The pronunciation lexicon distributed with the TIMIT database was adopted, with slight modifications. We have selected 46 different phones which are fully represented in our training vocabulary.

The most noticeable differences between the results shown in figures 1a and 2a are twofolded. First, the difference between native and non-native scores reduces slightly. On one hand the native scores decrease, which was expected. On the other hand, some non-native sub-sets were able to increase their scores.

The second difference relates to the advantages of using multiple mixtures which become noticeable when using sub-word modeling. This can be explained by the increase of the number of speech units available for training some of the sub-word models.

In the experiments reported in figure 2b, phone models were trained with each national group of speakers. As it was noticed for the whole-word recognisers, the use of specialised recognisers indicates a clear performance improvement over training with native speech. Again, none of this recognisers achieve the performance of the one tested with British speakers. However this difference became now as small as 3.4% for Portuguese speakers. We expected some performance degradation when changing from whole-word models to sub-word models, given the size of the vocabulary and the dimension of the training set. This degradation was found for British speakers, but was not so evident for non-native ones. In the case of the results obtained with German speakers, this evolution was even inverted.

The experiments presented in figure 2c concern models trained with a pool of all the training accents material. As expected, we obtained better results than when using only native speakers in the training set (figure 2a). However the improvements were not so marked as with whole-word models. Concerning the number of observation components, it becomes evident that the larger amount of training material made possible to train more components for an improved performance.

4. TRANSCRIPTION LATTICES

In the experiments reported above the same phonemic transcriptions were used for every accent. The use of a more flexible and accurate set of transcriptions is still dependent on the availability of human resources for manual labelling. Typical acoustic-phonetic decoders are based in ergodic topologies where all the sub-word models are represented. The transitions between these models are not retrained and the results are generally very poor with a high variability from utterance to utterance. These problems became clear very early for the researchers in continuous speech recognition [4]. Nowadays, there are a few methods that automatically derive the phonetic transcription in speech recognition tasks where many new words are expected [5]. Our goal here, is not the derivation of a single transcription but of a probabilistic model for the possible transcriptions, directly from a set of word repetitions.

4.1. Model definition

The automatic method we have used for deriving the probabilistic transcription lattices for a given word assumes the following input data:

- A number N_s (46) of models of the sub-word units $\lambda^{(i)} = (A^{(i)}, B^{(i)}, \Pi^{(i)})$.
- A number of speech signal repetitions from the word to be transcribed. This number should be as high as possible, so that every alternative path through the transcription lattice may represent a significant number of occurrences.
- The maximum number N_f of sub-word units for each word. We are currently working on alternatives that eliminate the need to specify this parameter. Pragmatical solutions can be easily derived based on the average duration of all the word repetitions or on the number of symbols of a standard orthographic transcription, but we also expect to derive formal solutions to this problem.

The method starts by the creation of an ordered series R_n $(n = 1, ..., N_f)$ of sets of sub-word models. Transitions between all models from set R_n to set R_{n+1} are allowed. In addition, there are initial and final non-emitting states which will be inserted in the previous series as R_0 and R_{N_f+1}. Thus, an HMM model $\lambda_{\mathrm{tr}} = (A_{\mathrm{tr}}, B_{\mathrm{tr}}, \Pi_{\mathrm{tr}})$, is obtained, which will be referred to as the transcription model. Each sub-word model in this transcription model has a defined temporal place to be selected, whereas in an ergodic network the sub-word models may occur at any time.

In our preliminary experiments with the transcription model, we do not intend to change the sub-word models themselves, which means that they are not reestimated. Since these models are repeated N_f times along the transcription model, they can be processed as fixed tied models. Thus, the only parameters affected by this contextual tagging are the transition probabilities between sub-word models. Hence, the transition matrix of the transcription model A_{tr} can be organised as a simpler matrix Γ, which we will call the inter-phone transition matrix. The sub-word models will be referred to as the macro-states of this matrix. The transitions are reestimated using the Baum-Welsh algorithm which maximises the global likelihood given the signal repetitions from the word to be transcribed.

Phone deletions and insertions are not allowed in the above description. However, the occurrence of these phenomena is one of the aspects which may be important for non-native speech recognition. In order to overcome this problem, transitions are allowed from every element of R_n to any element of R_m, provided that $m > n$. The transition probabilities are assumed to decrease with the topological/temporal distance $|m - n|$. These restrictions are imposed in the transition matrix of the initial model. By doing this, we expect to shape the reestimated models in a suitable form for deriving transcription lattices.

In order to create the matrix A_{tr} of the initial transcription model, we first build the matrix Γ according to the restrictions mentioned above, into which the transition matrixes A_i from each sub-word model are then inserted. The initial matrix Γ could also incorporate information such as provided from a pronunciation dictionary. However, this alternative was not yet explored.

After the reestimation process, there will be a new transition matrix A_{tr}. In order to simplify our representation for further processing, we must recover the corresponding matrix Γ.

4.2. Macro-states pruning

The inter-phone transition matrix Γ has a huge dimensionality which is not useful for deriving a transcription lattice, since it can not be manually checked neither used efficiently for recognition purposes. Most of the related macro-states are never (or only occasionally) visited during the reestimation process, which means they can be eliminated. Hence, a pruning process can reduce many of the $N_f * N_s$ macro-states. This pruning takes into account the state occupation counters at the end of the reestimation process. For each state, we sum the values in the corresponding counters obtained from all the repetitions.

In order to obtain an occupation measure for each macro-state, the counters from each state of the corresponding sub-word model are also summed (three in our case). This value is normalised by the number of repetitions available for training the word to be transcribed. Using this value, a threshold is tuned empirically, bellow which most of the macro-states are eliminated.

4.3. Results

The first transcription lattices obtained were estimated using only the speech material collected from the British speakers also used for training in previous experiments. If the pruning threshold is low enough, the best path can the found by visual inspection. Most of the time, the transcription thus obtained is very close, if not equal, to the phonemic transcription found in the pronunciation dictionary. The recognition results obtained with these lattices are summarised in table 1a, and can be compared with the best results of figures 1a and 2a, using whole-word and sub-word models. A slight increase of the overall performance was observed. For German speakers, the recognition score almost doubled.

In a second set of experiments, the British sub-word models were again used, with a Γ matrix trained with the each accent group of training speakers and tested with the corresponding test group. Comparing the results in table 1b with the scores previously obtained, improvements of nearly 20% in the recognition score can be found. If also the sub-word models are trained with specific accent material, the performance increases again (table 1c). However, the superiority of the transcription lattices is not so evident now.

5. CONCLUSIONS

Some of the conclusions from our previous work seemed to indicate the importance of having more detailed pronunciation transcriptions for each of the non-native accents present in our speech corpus. In order to derive a probabilistic transcription for each word, we have developed an approach where a finite-state-machine is initialised for each word. The corresponding states represent the sub-words and every possible sequence of sub-words is allowed. We have used a Baum-Welsh reestimation procedure in order to allow all utterances of the same word to contribute to the computation of state transitions. After this reestimation process, a pruning procedure eliminates the states with lower occupancy. The remaining full paths, together with their transition probabilities, thus provide relevant statistical information for building a transcription lattice. The most significant improvements were obtained when it was assumed that only native speech material was available for the sub-models training.

mix	da	de	en	es	it	pt	tot
(a) 3	69.7	40.8	95.3	57.5	41.6	62.8	59.0
(a) 6	65.0	27.4	92.5	50.8	42.3	65.0	54.8
(b) 3	76.3	58.6	95.3	79.5	63.0	78.0	74.1
(b) 6	74.1	57.5	92.5	77.8	63.3	80.5	73.3
(c) 3	83.1	61.4	95.3	82.5	70.5	87.9	79.1
(c) 6	88.0	72.6	92.5	84.9	78.1	90.9	83.2

Table 1. Percentage recognition scores using 3 and 6 observation mixtures. Training material: (a) British accent for sub-word models and transcription lattices; (b) British accent for sub-word models and specific accents for transcription lattices; (c) Specific accents for sub-word models and transcription lattices;

In the future, we plan to use these transcription lattices in our accent identification system. We will also investigate their potential for deriving improved phone-models and thus providing another dimension of recursive training for sub-word models and transcriptions.

REFERENCES

[1] J. Brousseau and S. Fox. Dialect-dependent speech recognisers for Canadian and European French. In *Proc. ICSLP*, pp. 1003–1006, Banff, 1992.

[2] L. Arslan and J. Hansen. Language accent classification in American English. In *Speech Communication*, vol. 188, pp. 353–367, 1996.

[3] C. Teixeira et al. Accent identification. In *Proc. ICSLP*, Philadelphia, 1996.

[4] F. Jelinek et al. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Information Theory*, vol. 21, pp. 250–256, 1975.

[5] R. Haeb-Umbach et al. Automatic transcription of unknown words in a speech recognition system. In *Proc. ICASSP*, pp. 840–843, Detroit, 1995.

[6] C. Teixeira and I. Trancoso. Word rejection using multiple sink models. In *Proc. ICSLP*, pp. 1443–1446, Banff, 1992.

[7] S. Young and P. Woodland. The use of state tying in continuous speech recognition. In *Proc. Eurospeech*, pp. 2203–6, Berlin, 1993.