OFF-LINE ACOUSTIC MODELLING OF NON-NATIVE ACCENTS

Silke Witt Steve Young
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ
United Kingdom
Email: {smw24,sjy}@eng.cam.ac.uk

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ABSTRACT

This paper introduces a family of three techniques to improve non-native speaker-independent recognition if the type of accent, i.e. the mother tongue of a non-native speaker is known. These techniques permit the computation of non-native models without requiring adaptation data, that is, they can be computed off-line. The improved recognition performance of these approaches is obtained by combining speaker-independent hidden Markov models of the target language, i.e. the language to be taught, and of the source language, i.e. the speaker's native language. All three combination techniques require a mapping between the two languages to define which source model state combines with each target model state. A method has been developed to derive such a mapping automatically. Recognition results are given for all three techniques applied to the two cases of Spanish accented British English and Japanese accented British English. The average baseline word error rate of 28.3% can be decreased to 22.9% for the first method, to 24.1% for the second and to 20.6% for the last method, which equals a relative improvement of 29% without adaptation.

1. INTRODUCTION

One of the major difficulties one encounters when trying to apply speech recognition to computer-assisted foreign language learning is the deterioration of speech recognition performance in the presence of heavily accented non-native speech. Generally speaking, current speaker independent recognition systems are known to perform considerably worse when recognising non-native speech. It has been shown by Chase [2] that this degradation is due to poor acoustic modelling. Similarly, Byrne et al.,[1], have demonstrated the need to improve the modelling of non-native speech. One way to improve non-native recognition is to retrain native speaker-independent models with non-native speech. In [3] retraining speaker independent models of American English with about 1000 sentences of Japanese accented English yielded a large improvement in recognition performance, bringing it up to the level of native recognition. However, this approach still requires a fairly large amount of non-native speech to retrain the models.

Given these problems, this paper introduces a set of techniques to improve non-native recognition performance without requiring adaptation data. The improvement is obtained solely through exploiting the knowledge of the mother tongue of the non-native speaker. These *accent prediction* methods require an additional model set of the native language and a mapping between the two languages in order to improve the recognition of foreign accented speech, but they do not require adaptation data. This approach is of practical importance, since it allows recognition systems to be used with non-native speakers (e.g. in a CALL application) without requiring an extensive enrolment phase.

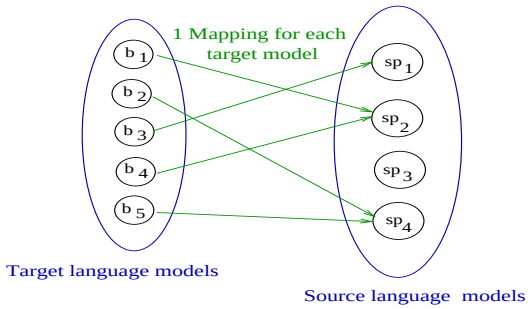

Figure 1. Mapping between target and source language models

All three techniques introduced here are based on the underlying idea that a non-native speaker — especially if he or she is just beginning to learn a target language — will substitute sounds of his or her mother-tongue for those foreign sounds he or she cannot produce. Such substitution patterns have been investigated extensively in the linguistic community, for example see [4]. Additionally, sounds of accented speech can be modelled by a mixture of sounds from the source and target language.

This paper is organised as follows: The next section describes the derivation of the necessary mapping between two languages. Such a mapping can either be based on phonetic literature or derived automatically using a small amount of non-native training data. In Section 3 the theoretical framework of three different methods of *accent prediction* is derived. Finally, these methods are experimentally evaluated in Section 4. The paper then concludes with a discussion of the experimental results.

2. MAPPING FROM TARGET TO SOURCE LANGUAGE MODELS

All model combination techniques which are presented in this paper are based on the idea that a given target model is likely to be substituted by a source model, or that the models of an accented model set can be considered to model a sound somewhere in-between a source and a target model. Consequently, one important task is to find a mapping between source and target language models, see also Figure 1. Such a mapping does not need to include all source models, nor is it a one-to-one mapping, because a source model can be a substitute sound for several target sounds.

Based on a suitable mapping, combining two models to a new model can improve the acoustic modelling. To be

Brit Phone	Literature	Corr based	Native Subs
ɑː	ɔː a_j	ae	a_j
ae	-	-	a_j
ʌ	a_j	ɑː	a_j
ɔː	a_j	ɒ	oo_j
aʊ	a + u	-	a_j
ə	?	ae	a_j
aɪ	-	-	a_j
b	b_j	-	p_j
ʧ	ʧ	-	ch_j
d	d_j	-	ch_j
ð	z d	z	t_j ch_j s_j
eə	-	-	ee_j
e	i eɪ	-	ee_j e_j
ɜː	ɑː aʊ	ɑː	aa_j
eɪ	eɪ	e	ee_j
f	h	-	s_j f_j
g	g_j	-	k_j
h	h_j	-	h_j
ɪə	-	ə	ii_j e_j
ɪ	i	e	i_j
iː	i	ɪ	ii_j
ʤ	ʤ	-	ch_j

Table 1. Typical error statistics for some sounds of Japanese accented English, using three difference source of knowledge, "-" denotes that no info was available, $-j$ denotes Japanese sounds

able to do so, it is necessary to learn which combination of models yields the optimal acoustic model of non-native sounds. This can be based on the knowledge of typical mispronunciations of foreign language students speaking the same mother tongue. In this paper three different approaches have been used in order to collect statistics about typical mispronunciations

1. Linguistics literature on pronunciation teaching containing listings of typical mistakes for a given source language, [5].
2. Corrected transcriptions of non-native speech from trained phoneticians, i.e. the comparison of the transcription of an utterance according to a pronunciation dictionary with the transcription of the same utterances corrected by a phonetician yields statistics about typical sound substitutions.
3. Automatic frame-level comparison of the results from the forced alignment of the target models with those from the alignment of a phone-loop using only source models. That source language phoneme which has been most often aligned with the location of a target phoneme yields the most likely substitution for this target phoneme.

While the first method requires manual derivation of the mapping and the second method requires hand-transcribed data, the third method is suitable for automatic derivation of a mapping for a given accent type provided that some training data (a few hundred sentences) of the required accent are available. In Table 1 likely substitutions for some English phonemes for each of the three methods are listed. As can be seen, the first two methods only yield substitutions for a subset of the phone inventory, whereas the third method provides a mapping for all phonemes. Based on such statistics as in Table 1 a suitable mapping can be derived by primarily using the third method. Then, the first two methods can be used to verify the mappings based on the third method. Also, if several source models are likely substitutions for one target model, only the most common one has been chosen

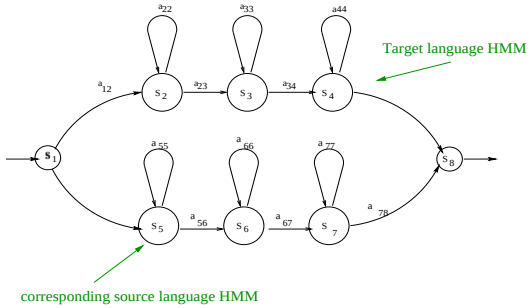

Figure 2. Bilingual HMM as a parallel combination of a source and a target language model

for a mapping. In [7] it has been shown that small variations in the mapping do not cause significant recognition performance degradation, thus these adaptation methods are not particularly sensitive to the exact mapping used.

3. THREE ACCENT PREDICTION METHODS

Given a mapping between a target and source language, there exist several ways of combining a target language model with its mapped source model. In this section three methods will be derived, all of them based on the same mapping, but using the following different model combination methods.

1. *Parallel Bilingual Modelling (PBM)*
2. *Linear Model Combination (LMC)*
3. *Model Merging (MM)*

3.1. Parallel Bilingual Modelling (PBM)

The most straightforward way of combining models of the source and target language into bilingual models is to combine each target model in parallel with its mapped source model. Creating such parallel models is equivalent to allowing substitutions of a source language model for each model in the target model set. In this example it is assumed that the original phone models have 3 emitting states plus non-emitting entry and exit states. Thus, in the new bilingual model the emitting states of the original models are in parallel sharing the entry and exit states. Figure 2 illustrates the topology of such a parallel model. Such a HMM combination requires to adjust the transition matrix. Given these parallel models, each phoneme in a word will be extended in a lattice for recognition to two phonemes in parallel, the correct phoneme and the possible substitution.

3.2. Linear Model Combination (LMC)

Assuming that the spectral space of a non-native speaker is somewhere in-between the space of a standard target language speaker and a standard source language speaker, suitable non-native models can be found through a linear combination of each target mixture component mean vector and its corresponding source mean vector. Define $\mathbf{B_s}$ as a diagonal matrix for state s in order to map from target mean μ_{T_s} to source mean μ_{S_s}. Thus, the $j-th$ diagonal element b_{js} represents the linear combination weight of the combination between source and target mean component. Then, each new mixture component mean is defined by the following construction:

$$\tilde{\mu_s} = \mathbf{B_s}(\mu_{S_s} - \mu_{T_s}) + \mu_{T_s} \quad (1)$$

Spkr	FL	PC	TS	MK	SS	Avg
Accent	Span	Span	Span	Jap	Jap	—
Baseline	20.3	29.4	26.5	19.1	45.8	28.2
PBM	15.5	23.1	21.1	16.5	38.2	22.9
rel Imp.	0.24	0.28	0.38	0.13	0.17	0.19

Table 2. WER Results and relative improvements for accent prediction with parallel bilingual models (PBM)

In [8] an algorithm has been presented to estimate the combination matrix $\mathbf{B_s}$ using adaptation data. However, in the case of off-line acoustic modelling, an *a-priori* estimate of $\mathbf{B_s}$ must used instead. Each diagonal element b_{js} of $\mathbf{B_s}$ represents a linear combination weight between a source mean vector element and a target one. Thus, each element b_{js} will be within the interval $[0, 1]$. Additional examination of the weights estimated with adaptation data suggests that models for improved non-native modelling typically have weights in the range of $0.0 < b_{apriori_j} < 0.5$. Therefore, average values in this range can be used for *a-priori* weights.

3.3. Model Merging

In the *Model Merging* (MM) algorithm each target model state is merged with its corresponding source model state by combining the two Gaussian mixtures into a new mixture with twice as many components as each original mixture. Again, the knowledge about the mapping between source and target language described in section 2 is exploited for this technique. Once the corresponding source model is known for each target model, the states of the models are merged sequentially, i.e. the first state of the target model is merged with the first state of the corresponding target model and so forth. Denote the target mixture as t and its mapped source mixture as s, then the new output probability density $b_i(\mathbf{o_t})$ of state i for the observation vector $\mathbf{o_t}$ is

$$b_i(\mathbf{o_t}) = \alpha \sum_{k=1}^{M_1} w_{sk} b_{sk}(\mathbf{o_t}) + (1-\alpha) \sum_{l=1}^{M_2} w_{il} b_{tl}(\mathbf{o_t}) \quad (2)$$

Here α denotes the weighting factor, w_{ik} the kth mixture component weight of state i. M_1 and M_2 denote the total number of mixture components in each mixture.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

All recognition experiments in this section use non-native speech from a specially recorded database, [8]. The test data consists of three speakers with a Hispanic and two speakers with a Japanese accent, each of them with 100 sentences. Three model sets of mixture Gaussian monophones for British English, Latin-American Spanish and Japanese have been build with the HTK Toolkit, [9]. The recognition system uses a 1000 word vocabulary and a word-pair grammar.

4.2. Parallel Bilingual Modelling (PBM)

First, the performance of PBM has been compared with a baseline using unadapted speaker-independent models of British English, see Table 2. By applying this technique a relative improvement of 19% over the baseline can be achieved. Comparing the recognition improvement of the individual speakers, it can be seen that in this case more performance gain is achieved for Spanish than for Japanese accented English. The requirements of this method are only a source-target mapping and a model set for the source language.

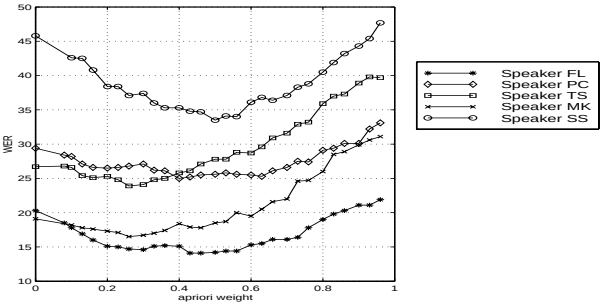

Figure 3. Word error rate dependency on the choice of *a-priori* weights. Weight 0.0 denotes the baseline error rate

Spkr	Base	$\text{LMC}_{jap,0.5}$	$\text{LMC}_{span,0.5}$
MK	19.1	18.5	23.3
SS	45.8	33.5	45.6
Avg.	32.5	26.0	34.5

Table 3. JAPANESE:Word error rate for accent prediction using models combined for a different accent.

4.3. Linear Model Combination (LMC)

In the following experiments the same constant weight factor has been used for all feature vector components b_{sj}. In Figure 3, this constant weight has been varied in the interval of $[0, 1]$. As can be seen the WER changes only slowly with changing weights. The implication for practical applications is that pre-programming a suitable weight will give improvements for most speakers of a given fluency level. However, the performance of this method depends on the individual speaker and therefore is less practical for off-line accent modelling than PBM. It is interesting to note that the optimal *a-priori* weight for a given speaker can be considered as a measure of his or her speaking fluency. For example, speaker SS in Figure 3 is significantly less fluent than speaker TS. If the optimal weight is small, the newly combined models will mostly contain components of the target language model and it can be concluded that the student is more fluent than someone whose optimal weight is much larger.

Another experiment serves as a check on whether indeed the inclusion of phonetic knowledge contributes to the improved recognition rather than simply the increased amount of information obtained by doubling the amount of available parameters. A model set of a different language other than the source language has been combined with the target language models. Thus, the Spanish and English models have been combined with the *a-priori* weight $p = 0.5$ to recognise Japanese accented speech. The results are shown in Table 3. Using LMC predicted models based on Spanish as the source language yields 6% worse recognition relative to the baseline, whereas the performance of LMC with Japanese source models increases by 20%. This is a clear indication that the use of knowledge about the target language does contribute to the improvements achieved with the accent prediction techniques presented in this paper.

4.4. Model Merging (MM)

The remaining prediction method to be tested is Model Merging. In this case, the only variable which has to be estimated is the weighting factor α. In the following experiment this factor has been varied between $\alpha = 0.0$, i.e. using only the target models and $\alpha = 1.0$, i.e. using only the source models. In Figure 4 the WER is shown as

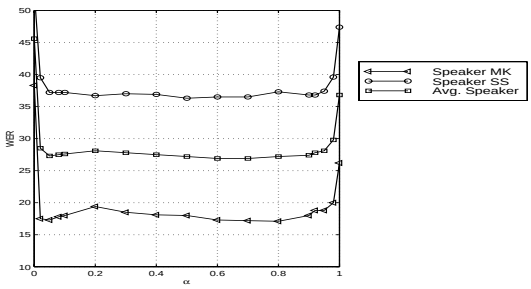

Figure 4. WER for accent prediction with dependency on merging weight α. $\alpha = 1.0$ denotes target models only

Spkr	FL	PC	TS	MK	SS	Avg
Accent	Span	Span	Span	Jap	Jap	—
Baseline	20.3	29.4	26.5	19.1	45.8	28.2
MM	12.1	19.6	18.9	17.5	34.9	20.6

Table 4. WER Results for accent prediction with model merging (MM) in comparison with the baseline

a function of the merging coefficient α. It can be seen that the recognition accuracy does not vary significantly for any $0.1 \leq \alpha \leq 0.9$. The fact that the performance doesn't depend on α makes MM a robust method for applications.

From Table 4 it can be seen that accent prediction based on model merging can reduce the word error rate from $WER = 28.3\%$ for the baseline to $WER = 20.6$. This equals a relative improvement of 27.2%.

In Table 5 the results of maximum likelihood linear regression (MLLR) adaptation with a full global transformation matrix and 5 iterations, [6], are compared with MM. As shown, MM performs better than MLLR with 6 adaptation by 5% relative. If significant amounts of adaptation data are available, MLLR outperforms accent prediction with MM. However, if adaptation data are used to adapt the mixture component means and to re-estimate the mixture component weights, then MM performs significantly better than MLLR for less adaptation material. These results show that the use of pre-combined models as initialisation models for MLLR-based adaptation can yield improvements over the adaptation of the target language models alone.

5. CONCLUSIONS

Three different techniques have been presented which improve the acoustic modelling of foreign accented speech without requiring adaptation material. The improvements are gained solely through exploiting the knowledge of possible substitutions errors expressed in the form of a model mapping between a given language pair and through using an additional model set of the source language. A performance summary is given in Table 6. The most successful method, Model Merging, yields a relative recognition improvement of 27% over the speaker-independent baseline system. These results indicate that using additional information about a speaker's mother-tongue can improve recognition performance without requiring on-line adaptation. Also, if only limited adaptation data are available, MM can still outperform MLLR for this type of heavily accented speech.

Spkr	Base	MM	MLLR_6	MLLR_{24}	MM_6
Avg.	28.3	20.6	21.7	19.2	16.8

Table 5. Comparison of MM and MLLR adaptation. MLLR_6 and MM_6 use 6 adaptation sentences, MLLR_{24} uses 24

Spkr	Base	PBM	$\text{LMC}_{0.3/0.46}$	MM
Avg.	28.3	22.9	24.1	20.6

Table 6. WER summary for accent prediction using bilingual models

6. ACKNOWLEDGEMENTS

Silke Witt is funded by an EPSRC advanced studentship and a Marie Curie Research Fellowship of the European Union.

REFERENCES

[1] W. Byrne, Knodt E., S. Khudanpur, and J. Bernstein. Is automatic speech recognition ready for non-native speech? a data-collection effort and initial experiments in modeling conversational hispanic english. In *Proceedings STiLL*, pages 37–40, Marholmen, Sweden, 1998.

[2] L.L. Chase. *Error-responsive Feedback Mechanisms for Speech recognisers.* PhD thesis, Carnegie Mellon University, Pittsburgh,USA, 1997.

[3] F. Ehsani. Ntt-data japanese-english atc asr system description. Technical report, Entropic, Inc., 1996.

[4] J.E. Flege. Production and perception of a novel, second-language phonetic contrast. *J. Acoust. Soc. Am.*, 93(3):1589–1608, March 1993.

[5] J. Kenworthy. *Teaching English Pronunciation.* Longman, 1987.

[6] C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Department, Cambridge, U.K., June 1994.

[7] S.M. Witt. *Use of Speech Recognition in Computer-assisted Language Learning.* PhD thesis, Cambridge University Engineering Department, 1999.

[8] S.M. Witt and S.J. Young. Bilingual model combination for non-native speech recognition. In *Proc. Institute of Acoustics Conference on Speech and Hearing*, 1998.

[9] S. J. Young, J. Odell, D. Ollason, and P. C. Woodland. *The HTK Book.* Entropic Cambridge Research Laboratory, 1996.