Proceedings of the Institute of Acoustics

BILINGUAL MODEL COMBINATION FOR NON-NATIVE SPEECH RECOGNITION

S M Witt (1), S J Young (1)

(1) University of Cambridge, Engineering Department, Cambridge CB2 1PZ, UK

1. INTRODUCTION

Current speaker independent recognition systems are known to perform considerably worse when recognising non-native speech, [2, 3]. On average, such speech has much slower speaking rate (up to 20% in our data) and contains hesitations as well as mispronunciations. Also, the intonation pattern can differ considerably from native speech. Given sufficient foreign-accented training data, hidden Markov models for a specific foreign accent could be retrained using Baum-Welch re-estimation. However, generally such data are sparse. The objective of the work presented here is to rapidly adapt to a specific accent using only as little adaptation data as possible. Such adaptation to non-native speech can be regarded as a special case of accent adaptation. However, unlike other adaptation schemes which are based on matrix transformations from a speaker independent system to the accent specific acoustic space, the approach presented here is based on finding the optimal acoustic space through combining mean vectors of the hidden Markov models of the target language and the source language, i.e. a speaker's native language. It is hoped incorporating information from both languages provides more direction information in the acoustic space.

The following section introduces the theoretical framework for the new adaptation technique. Section 3 discusses how to determine mappings between the phoneme sets of two languages. Finally, the new adaptation scheme is tested on a non-native database specifically recorded and transcribed for research on the speech of foreign language students, [7]. All experiments have been based on British English as the target language and Spanish as the source language.

2. BILINGUAL MODEL COMBINATION

For a recognition system based on continuous density mixture Gaussian hidden Markov models the following definitions are required: Consider a model set of the target language, M_T, which contains Q_T models and a set of the source language, M_S, which contains Q_S models. Let a continuous density HMM may have N states with mean μ_i for state i. The dimension of the observation space is J.

The output probability of an observation vector for a mixture component given the state is defined as

$$b_{ik}(\mathbf{o}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{C}_{ik} \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{o}-\mu_{ik})' \mathbf{C}_{ik}^{-1} (\mathbf{o}-\mu_{ik})} \tag{1}$$

where the M mixture component densities are combined to give the state probability density:

$$b_i(\mathbf{o}) = \sum_{k=1}^{M} \mathbf{w}_{ik} \mathbf{b}_{ik}(\mathbf{o}) \tag{2}$$

The goal of model combination is to find the re-estimated mixture means $\hat{\mu}$, through a weighted linear combination of target language mixture means and source language mixture means. To do so, define $\mathbf{B}_s$ as a $J \times J$ dimensional diagonal matrix for state s which

maps from target mean μ_{T_s} to source mean μ_{S_s}.

$$\bar{\mu}_s = \mathbf{B}_s(\mu_{S_s} - \mu_{T_s}) + \mu_{T_s} \tag{3}$$

Thus, the $j-th$ diagonal element $b_s(j)$ represents the linear combination weight for the source and target means component.

For simplicity, the re-estimation formulae will firstly be derived for the case of single component Gaussians. Based on these equations, the extension to multiple component mixtures can be derived in a straightforward manner.

2.1 SINGLE GAUSSIAN DISTRIBUTIONS

The mean vector of a Gaussian distribution is adapted by a linear combination of a mean vector of the source language and a mean vector of the target language. For now, it is assumed the mapping of target means to source means is known.

Let $\mathcal{F}(O,\theta|\lambda)$ be the likelihood of generating the observed speech frames while following the state sequence $\theta = (\theta_0\theta_1,\ldots\theta_T)$ given the set model parameters λ. Then it is convenient to define an auxiliary function $Q(\lambda,\bar{\lambda})$

$$Q(\lambda,\bar{\lambda}) = \sum_{\theta\in\Theta} \mathcal{F}(O,\theta|\lambda) log(\mathcal{F}(O,\theta|\bar{\lambda})) \tag{4}$$

Estimating the model parameters is now based on iteratively maximising this auxiliary function by improved parameters and forming a new auxiliary function with the improved parameters.

A re-estimation expression for the state means is found through differentiating $Q(\lambda,\bar{\lambda})$ with respect to $\bar{\mu}$ using equation (3). To do so, define the probability of state occupancy as

$$\gamma_s(t) = \frac{1}{\mathcal{F}(O|\lambda)} \sum_{\theta\in\Theta} \mathcal{F}(O,\theta_t = s|\lambda) \tag{5}$$

To estimate the matrix $\mathbf{B}_s$, it is necessary to differentiate $Q(\lambda|\bar{\lambda})$ with respect to $\mathbf{B}_s$ and equate it to zero:

$$\frac{dQ(\lambda,\bar{\lambda})}{d\mathbf{B}_s} = \frac{d}{d\mathbf{B}_s}\left[\sum_{j=1}^{J} c_j(o_{t,j} - \bar{\mu_{s,j}})^2\right] \tag{6}$$

Substituting 3 for $\bar{\mu}$ gives the derivative of this with respect to each diagonal element b_j of $\mathbf{B}$

$$\frac{\delta h}{\delta b_j} = c_j\left[o_{t,j} - b_j(\mu_{S,j} - \mu_{T,j}) + \mu_{T,j}\right](\mu_{S,j} - \mu_{T,j}) \tag{7}$$

This gives a set of N equation for the N diagonal elements of $\mathbf{B}_s$. The equation for the i-th element is

$$\sum_{t=1}^{T} \gamma_s(t)\left[o_{t,j} - b_j\left\{\mu_{S,j} - \mu_{T,j}\right\} - \mu_{T,j}\right]\left\{\mu_{S,j} - \mu_{T,j}\right\} c_j = 0$$

yielding

$$b_j = \frac{\sum_{t=1}^{T} \gamma_s(t)\left[o_{t,j} - \mu_{T,j}\right]}{(\mu_{S,j} - \mu_{T,j})} \tag{8}$$

2.2 TIED COMBINATION MATRICES

Given the problem of data sparseness, it is desirable to extend the above derivation to the case of tied combination matrices. If a $\mathbf{B}_s$ matrix is shared by R states $\{s_1, s_2 \ldots s_R\}$ the derivative becomes

$$\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)\left[o_{t,j} - b_j\left\{\mu_{S_r,j} - \mu_{T_r,j}\right\} - \mu_{T_r,j}\right]\left\{\mu_{S_r,j} - \mu_{T_r,j}\right\}c_{r_j} \quad = \quad 0$$

Solving for b_j yields

$$b_j = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)c_{r_j}\left[o_{t,j} - \mu_{T_r,j}\right](\mu_{S_r,j} - \mu_{T_r,j})}{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)c_{r_j}(\mu_{S_r,j}) - \mu_{T_r,j})^2} \tag{9}$$

The extension of the above derivations to mixture Gaussians is straightforward, see [5]. The only mathematics which changes in equation (9) are the indices.

3. MAPPING

The mapping between each target mean and source mean is crucial for the performance of the model combination technique. Three mapping criteria have been tested to define the relationship between parameters of the source and target model sets.

3.1. MIXTURE MAPPING BASED ON ACOUSTIC DISTANCE

A straightforward mapping for each mixture component of the target system is to determine the closest mixture component of the source system using an acoustic distance measure. In this work two such measures have been compared: The standard Euclidean distance and a divergence measure, which measures the separability of two Gaussians $\mathcal{N}(\mu_A, \Sigma_A)$ and $\mathcal{N}(\mu_B, \Sigma_B)$

$$D_{div} = \frac{1}{2}trace\left(\frac{\Sigma_A}{\Sigma_B} - I\right) + \frac{1}{2}(\mu_A - \mu_B)^T\frac{(\mu_A - \mu_B)}{\Sigma_B} + \frac{1}{2}ln\frac{|\Sigma_B|}{|\Sigma_A|} \tag{10}$$

3.2. STATE MAPPING BASED ON ACOUSTIC DISTANCE

The state mapping method maps each state to the closest state of the source model set. The state distance measure in equation (11) is employed to find for each state of the target language the nearest state of the source language.

$$d(i,j) = -\frac{1}{S}\sum_{s=1}^{S}\frac{1}{M_S}\sum_{m=1}^{M_S} log[b_{j_s}(\mu_{i_{sm}})] + log[b_{i_s}(\mu_{j_{sm}})] \tag{11}$$

Some examples of a state mapping between Spanish as the source language and British English as the target language are listed in Table 1. Within each target state, all mixture components are mapped to the closest mixture component mean of the corresponding source state using the divergence distance measure.

B Phone State	Sp Phone State	B Phone State	Sp Phone State
æ 2	a_s 3	æ 4	a_s 4
ʌ 2	o_s 2	ʌ 3	a_s 4
ə 2	e_s 3	ə 4	n_s 3
ð 4	$ð_s$ 4	h 3	j_s 3

Table 1: State mapping from British English models to Spanish models using the state distance measure. The subscript 's' denotes a Spanish model. IPA symbols are used to describe the phonemes.

3.3. MODEL MAPPING BASED ON PHONETIC KNOWLEDGE

The two previous mappings are solely based on acoustic distances and do not include phonetic knowledge about the likely mispronunciations of a language student. On the other hand, this third model combination technique is 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. Thus, a combination of these two models yields improved acoustic modelling. To do this, it is necessary to understand which combination of models represents the non-native speech. A first understanding of the relationship between the two model sets can be obtained through statistics about which phoneme is likely to be substituted because of pronunciation errors. Information about typical mispronunciation for foreign language students speaking the same mother tongue can be collected through the following three different approaches:

1. Linguistics literature on pronunciation teaching containing listings of typical mistakes for a given source language, [4].

2. Corrected transcriptions of non-native speech from trained phoneticians.

3. 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. This yields likely substitutions of source models for individual target models.

In Table 2 typical errors are shown for some British English phonemes based on these three methods. A full listing can be found in [6]. For those fields containing a dash either no typical pronunciation error was mentioned in the literature or the amount of corrections made by human judges is not statistically significant. Comparing the predicted errors, these three different sources show fairly strong agreement. In the case when a mapping for a new language pair is needed, the necessary information can be found with the help of one or any combination of these methods. Take for example the phoneme "b". In [4] it is noted that people with a Spanish accent tend to substitute this sound with their native "b" sound, which is somewhere between the English "b" and "v". In Table 2 the Spanish "b" is listed as an error mentioned in the literature (first column), the human judges often marked the "b" with "bv", their made-up symbol for this in-between sound (second column) and finally in the alignment with Spanish models, the Spanish "b" is most often aligned with the English "b". Applying the information in Table 2, the most likely substitution is found for each target model and used to map from the source to the target model. The states are mapped sequentially, i.e. the first state of the target model maps to the first state of

British Phone	1. Phon. Knowledge	2. Human Corr.	3. Native Subs
ɑː	-	æ	a_s
æ	e	ɑː, ʌ	a_s
ah	ə, ɒ	ə, ɒ	a_s, o_s
ɔː	-	ɒ, r	o_s
aʊ	-	-	o_s
ə	-	æ, e, oh, uː	a_s, e_s, r_s
aɪ	-	I	a_s
b	β	bv	β_s
ʧ	-	-	$ʧ_s$
d	ð	ð, del	d_s, t_s
ð	-	d	d_s, t_s
eə		ɜː	e_s, a_s

Table 2: Probable substitutions for British phonemes based on three different knowledge sources: 1) Phonetic knowledge, 2) Based on statistics of teachers' corrections, 3) Most likely Spanish phonemes when aligning the data with Spanish models only. IPA symbols are used to describe the phonemes.

the source model. Within each state the closest source mixture component is found for each target mixture component using the divergence measure.

4. ACCENT PREDICTION

In Figure 1 the coefficients are shown for two non-native and one native speaker. As expected, the coefficients for non-native speakers are on average greater than zero, in contrast to the native speaker whose coefficients vary around zero. This observation leads to the approach of finding a set of weights to combine the two model sets *a-priori* to any further adaptation. These a-priori weights are chosen considering the fact that the combination weights for non-native speakers tend to be in the interval $[0, 0.2]$. Thus, weights $b_i = 0.1$ are used. Secondly, since foreign accents cause changes particularly in the second and higher formats of accented speech [1], some of the weights representing the lowest mel-frequency cepstral coefficients, are set to zero. Using such a set of predefined *a-priori* weights yields a method of accent prediction. This is because a model set for a particular non-native accent can be estimated off line and independent of any adaptation data, by using the model combination technique with these *a-priori* weights.

Like most maximum likelihood estimators, the one discussed in this paper is shown to find local, but not global, maxima. Starting the estimation process with a different initialisation of the models might lead to the estimation of different local optima and thus to improved models. Therefore, non-native speech adaptation can be improved by using models based on model combination with predefined weights as the basis for the adaptation process. This can be achieved using either using Maximum Likelihood Linear Regression (MLLR) or model combination.

5. EXPERIMENTAL RESULTS

The data used for all experiments presented in this section are part of a non-native database, see [7]. This database contains three speakers speaking Latin-American Spanish as their mother-tongue. Each speaker set consists of 120 sentences of read speech, with a 1000 word vocabulary. As source models,

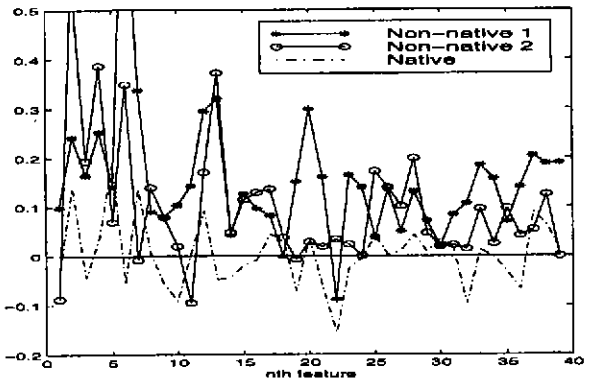

Figure 1: Example of model combination weights for two non-native speakers and one native speaker

Spkr	Base	MLLR	MC_m	MC_s	MC_A	MC_B
FL	20.3	20.5	19.1	19.0	17.9	16.0
PC	29.4	28.4	30.1	30.3	28.4	26.8
TS	26.7	26.3	26.0	26.0	24.7	25.2
Avg.	25.5	25.0	25.1	25.1	23.7	22.7

Table 3: Word error rate for baseline, standard MLLR, mixture mapping model combination (MC_m), state mapping (MC_s, model mapping using Eucledian distance (MC_A and model mapping using the divergence measure (MC_B, all experiments using a global transform and 5 adaptation sentences

mixture Gaussian monophones trained on Standard British English were used. The target models were trained on Latin-American Spanish. The HTK Toolkit, [8] was used for both the model training and for the recognition experiments. As language model a word-pair grammar based on textbooks for learners of English as a second language was used. This is because the text spoken by the language students had also been taken from these textbooks. In all experiments, the results for an equivalent set-up of MLLR were measured too, in order to have a comparison benchmark. Because rapid adaptation with only a few sentences is of interest, only 5 adaptation sentences and a global transform, tying all models into one class, were employed.

In Table 3 the results for MLLR are contrasted with the results for model combination based on four different mappings. Both the mixture mapping, MC_m, and the state mapping, MC_s, are calculated using the divergence distance measure. The model mapping MC_A uses the Euclidean distance to map between the mixture components within a state, whereas the model mapping MC_B uses the divergence measure. The influence of the choice of mapping is easily seen. Whereas mixture and state mappings perform similar to MLLR, both model mappings yield significant improvements. It can also be seen

Spkr	Base	MC_A (w1)	MC_A (w2)	MC_A (w3)	MC_B	A-MLLR	A-MC
FL	20.3	19.7	18.8	17.8	16.9	20.3	16.4
PC	29.4	28.4	28.0	28.2	27.1	26.2	27.0
TS	26.5	26.3	26.0	26.4	25.4	26.4	25.1
Avg.	25.5	24.8	24.3	24.1	23.1	24.3	22.8

Table 4: Word error rate for accent prediction. MC_A: Model-level mapping, using the divergence measure and a-priori weight=0.1 (w1: first 8 coeff to 0.0, w2: first 5, w3: none), MC_B same model mapping, but a-priori weight 0.13, A-MLLR : MLLR adaptation of predicted models, A-MC: MC adaptation of predicted models.

that the divergence measure, which incorporates variances as opposed to the Euclidean distance measure which only uses the means, is the more effective distance measure. For the best mapping, MC_B, a relative improvement of 11% compared to the baseline and of 9% compared to MLLR was measured. These results indicate that knowledge about the mother tongue of a non-native speaker (and thus of the likely phoneme substitutions) improves the acoustic modelling.

Next, results for accent prediction are displayed in Table 4. The number of zero weights is varied from 8 (w1) to none (w3). Overall, the best results are achieved for no zero weights. In all cases an improvement over the baseline recognition rate occurs. Varying the weights' value yields the highest improvement of 23.1% for $b_i = 0.13$. This is a relative improvement of 9.4% over the baseline. Thus this accent prediction approach can improve recognition without requiring adaptation data. However, the use of pre-combined models as initialisation models for either adaptation scheme did not yield improvements over the adaptation of the target language models.

6. CONCLUSIONS

In this paper we presented a novel adaptation technique for non-native data based on a linear combination of each HMM mixture means from the target language model set with a mixture mean from the source language model set.. This scheme has low computational requirements and is shown to be effective for even little adaptation data. A relativ adaptation improvement of 11% over the baseline can be achieved with as few as 5 adaptation sentences. Also, word error rate decrease by 9% relative to the standard MLLR algorithm which is used as a comparison baseline.

The model combination technique requires a mapping between the source and target language model sets. Since this mapping is crucial for the performance of the algorithm presented here, several mapping approaches have been discussed. Experimental results show that a mapping which incorporates phonetic knowledge about likely mispronunciations of a non-native speaker yields the largest performance increase.

Thirdly, a method of accent prediction is introduced. The use of *a-priori* defined weights enables off-line pre-computation of a model set for a given accent. Recognition performance of such accent models can decrease the baseline word error rate from 25.5% to 23.1%. This means that using additional information about a speaker's mother-tongue can improve recognition performance significantly without requiring on-line adaptation.

Further work will investigate additional modifications to this adaptation approach, especially with regard to using triphones instead of mixture Gaussians. Also, the choice of*a-priori* weights for accent prediction needs further work.

Finally, further investigation on how such model combination techniques can be used within the framework of computer-assisted language learning is necessary, because recognition of heavily accented speech is required in such systems.

7. REFERENCES

[1] L. Arslan and J.H.L. Hansen. Frequency characteristics of foreign accented speech. In *ICASSP '97*, Munich, Germany, April 1997.

[2] 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.

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

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

[5] 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.

[6] S.M. Witt. A non-native database for foreign language learning. internal notes.

[7] S.M. Witt and S.J. Young. Performance measures for phone-level pronunciation teaching in call. In *STiLL:Speech Technology in Language Learning*, pages 99–102, Marholmen, Sweden, May 1998. ESCA.

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