Estimation of models for non-native speech in computer-assisted language learning based on linear model combination

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

Abstract

This paper investigates how to improve the acoustic modelling of non-native speech. For this purpose we present an adaptation technique to combine hidden Markov models of the source and the target language of a foreign language student. Such model combination requires a mapping of the mean vectors from target to source language. Therefore, three different mapping approaches, based on either phonetic knowledge and/or acoustical distance measures have been tested. The performance of this model combination method and several variations of it has been measured and compared with standard MLLR adaptation. For the baseline model combination small improvements of recognition accuracy compared to the results based on applying MLLR were obtained. Furthermore, slight improvements were found when using an a-priori approach, where the models were combined with predefined weights before applying any of the adaptation techniques.

1. Introduction

Current speaker independent recognition system are known to perform considerably worse when recognising non-native speech. Chase showed that such performance deterioration is due to bad acoustic modelling, [3]. Similarly, initial investigations about recognition characteristics of non-native speech made by Byrne et al.,[2] demonstrated the need to improve the modelling of non-native speech. In this paper we present a technique to adapt to non-native speech which deploys the additional information, which is given if the kind of accent, i.e. the mother-tongue of the speaker is known. This technique is based on linearly combining each mean vector of a model from the target language with the mean vectors of a model of the source language. The combination weights can be estimated by applying re-estimation formulas similar to those used in the MLLR adaptation algorithm, [5].

Unlike other adaptation schemes which are based on a matrix transformation from a speaker independent system to the accent specific acoustic space, the approach presented here constrains the search to the space between the model sets of the two languages involved. Incorporating information from both languages is hoped to provide more direction in the acoustic space towards the location of better models for non-native speech and potentially allow faster adaptation on small amounts of data.

This work represents parts of an ongoing project to investigate the use of automatic speech recognition in computer assisted language learning (CALL), [6]. The techniques to score pronunciation which have been developed so far within this project, have been based on the assessment of the pronunciation of read speech by a student. In a next step it is desired to recognise a student's speech as spoken in a dialog with the computer. In such a setup it is necessary to model non-native speech. Compared with native speech, non-native speech is characterised by different spectral characteristics especially in the higher formants, see also [1]. By modifying speaker independent models of the target language with components of the source language, it is hoped to account for these spectral differences.

One of the additional characteristics of the corpus of non-native speech spoken by students of English used for the experiments presented here is that the students speak haltingly. They also make pronunciation errors (about 20% of transcriptions has been error marked by phoneticians) and the speech rate is on average reduced by a factor of 1.2.

In the next section the theoretical framework of the bilingual adaptation algorithm will be derived. Additionally, two modifications of the basic adaptation approach will be presented as well. This section is followed by a discussion of three different mapping approaches between the source and target language will be discussed. In section 4 the experimental results will be presented.

2. Derivation of Linear Model Combination

Let M_T be a model set of the target language containing Q_T models, and M_S a model set of the source language with Q_S models. Assume a continuous density multiple mixture HMM with N states, transition probabilities a_{ij}, where the output probability of the ith state, b_i for a speech frame vector $\mathbf{o}$ is given as

$$b_i(\mathbf{o}) = \sum_{k=1}^{M} w_{ik} b_{ik}(\mathbf{o}) \tag{1}$$

with the output probability of each mixture component given as

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

where μ_{ik} denotes the mean of mixture component k (vector of length n), w_{ik} the mixture weight and C_{ik} the $n \times n$ covariance matrix.

2.1. Single Mixture Gaussians

Firstly, the re-estimation expressions will be derived for the case of single mixture HMMs. Assume a mapping of each target mean to a source mean. Then a new mean can be estimated as

$$\tilde{\mu_s} = \mathbf{B_s}(\mu_{S_s} - \mu_{T_s}) + \mu_{T_s} \tag{3}$$

where $\mathbf{B_s}$ is defined 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 jth diagonal element $b_{s,j}$ represents a linear combination weight for the respective source and target mean vector elements. The parameters of any linear model space transformation can be found through application of the EM algorithm. The method presented here represents a modification of the MLLR adaptation algorithm, see [5]. The auxiliary function which has to be minimised can be written as

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

where λ denotes the current set of model parameters and $\bar{\lambda}$ a re-estimated set of parameters. The likelihood of generating the observed speech frames in the state sequence θ is

$$\mathcal{F}(O, \theta|\lambda) = a_{\theta_T N} \prod_{t=1}^{T} a_{\theta_{t-1}\theta} b_{\theta_t}(\mathbf{o_t}) \tag{5}$$

To estimate the matrix $\mathbf{B_s}$, it is necessary to differentiate $Q(\lambda, \bar{\lambda})$ with respect to $\mathbf{B_s}$ using equation 3 and equate it to zero:

$$\frac{dQ(\lambda, \bar{\lambda})}{d\mathbf{B_s}} = 0 \tag{6}$$

Solving this derivative for the j-th element of $\mathbf{B_s}$ yields (index j indicates a vector component):

$$b_{s,j} = \frac{\sum_{t=1}^{T} \gamma_s(t) \left[o_{t,j} - \mu_{T_s,j}\right]}{\sum_{t=1}^{T} \gamma_s(t)(\mu_{S_s,j} - \mu_{T_s,j})} \tag{7}$$

where $\gamma_s(t)$ the probability of state occupancy at time t.

2.2. Tying of Model Means into Regression Classes

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 solution for b_j becomes

$$b_j = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{s_r}(t) c_{r_j} \left[o_{t,j} - \mu_{T_r,j}\right] (\mu_{S_r,j} - \mu_{T_r,j})}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{s_r}(t) c_{r_j} (\mu_{S_r,j} - \mu_{T_r,j})^2} \tag{8}$$

The above derivation of the estimation of the combination matrix can be extended to the case of multiple mixtures in a straightforward way, as they can be pictured as multiple weighted states. For an equivalent derivation see [5]. All that changes in equation (8) are the indices.

2.3. Bilingual Model Alignment

Similar to MLLR, this model combination technique requires alignment of the transcriptions of the adaptation sentences with the speech data. An approach to improve modelling of non-native speech could be to include the model set of the speaker's mother tongue in the alignment stage. Using a mapping from each target model to one source model, a recognition network was built consisting of a sequence of target models according to the transcriptions. The mapping source model was put in parallel with its respective target model. Thus, transcriptions consisting of phonemes of both languages were calculated and used for the re-estimation process.

In the case that a phoneme from the source language is given in the transcriptions, the re-estimated mean is defined as:

$$\tilde{\mu_s} = (\mathbf{I} - \mathbf{B_s})(\mu_{T_s} - \mu_{S_s}) + \mu_{S_s} \tag{9}$$

otherwise the old definition remains:

$$\tilde{\mu_s} = \mathbf{B_s}(\mu_{S_s} - \mu_{T_s}) + \mu_{T_s} \tag{10}$$

Thus, for part of the training data the new mean is estimated based on accumulated statistics for the initial target mean, and for the other part of data, the new mean is estimated starting out from the source mean. Define T_1 as the number of frames associated with state s using the British model, and T_2 the number of frames using the Spanish model. With this approach the auxiliary function rewrites as

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{N} Q_{a_i}[\lambda, \{\bar{a}_{ij}\}_{j=1}^{N}] +$$

$$\sum_{k \in S} \mathcal{F}(O|\lambda) \left[\sum_{t=1}^{T_1} \gamma_{s_T}(t) log \bar{b_{s_T}}(\mathbf{o}_t) + \sum_{t=1}^{T_2} \gamma_{s_S}(t) log \bar{b_{s_S}}(\mathbf{o}_t) \right] \tag{11}$$

Solving for $b_{s,j}$ yields

$$b_{s,j} = \frac{\left[\sum_{t=1}^{T_1} \Delta_{\mu_{T_s}} + \sum_{t=1}^{T_2} \Delta_{\mu_{S_s}} - \sum_{t=1}^{T_2} \Delta_{\mu_{T_s},\mu_{S_s}})\right] \delta_{T_s,S_s}}{\left[\sum_{t=1}^{T_1} \gamma_{s_{T_s}}(t) c_{T_s,j} + \sum_{t=1}^{T_2} \gamma_{s_{S_s}}(t) c_{S_s,j}\right] \delta^2_{T_s,S_s}} \tag{12}$$

with

$$\begin{aligned} \delta_{T,S} &= (\mu_{S,j} - \mu_{T,j}) \\ \Delta_{\mu_T} &= \gamma_{s_T}(t) c_{T,j} \{o_{t,j} - \mu_{T,j}\} \\ \Delta_{\mu_S} &= \gamma_{s_S}(t) c_{S,j} \{o_{t,j} - \mu_{S,j}\} \\ \Delta_{\mu_T,\mu_S} &= \gamma_{s_S}(t) c_{S,j} \delta_{T,S} \end{aligned}$$

2.4. A-priori Combination

From Figure 2 it is apparent that an increase in the number of adaptation sentences does not significantly increase the recognition performance. One possible explanation is that inaccurate modelling by the target models causes a large number of alignment errors which makes MLLR adaptation ineffective. Like most maximum likelihood estimators, the ones discussed in this paper have been shown to find local maxima, but not global ones. Starting the estimation process at a different location might yield different local optima and thus improved models. Therefore, we propose a third model combination approach which combines source and target models using a-priori weights before executing re-estimation. For example, examination of the weights shown in Figure 1 suggests that models with improved non-native modelling are likely to have combination weights in the range of $0.0 - 0.2$.

3. Mapping from Target to Source Means

The model combination technique requires a mapping of each target model mean vector with a mean vector of the source language. This mapping can be based on acoustical distance measures and/or phonological knowledge. We experimented with three different approaches

1. **Mixture-level:** For each target mean the source mean with minimum Euclidean distance was found. This mapping is based solely on acoustic distances and as such disregards any connectivity between mixture means of a state and a model.
2. **State-level:**: This mapping approach moves up to state level and calculates the closest source state for each target state, using the following state distance measure:

$$d(i,j) = -\frac{1}{S}\sum_{s=1}^{S}\frac{1}{M_S}\sum_{m=1}^{M_S} log[b_{js}(\mu_{ism})] + log[b_{is}(\mu_{jsm})] \tag{13}$$

   The mixture component for each state were then mapped using to Euclidean distances as in (1).
3. **Model-level:** In order to find out which source models are likely to be substituted by non-native speakers for target models, we compared the forced alignment results of the target models with the alignments results of a phone-loop with source models. For instance, let British English be the target language and Spanish be the source language. Then some examples of those phonemes which are likely to be substituted are the Spanish 'b/v' sound for both the English 'b' or 'v' or the Spanish 'rr' for the English 'r'. Each target model was mapped to that source model which was most often aligned at the same time. Given this mapping on model level, the states of the model pair where mapped in order, i.e. state 2 to state 2 etc, and the mixtures were mapped using Euclidean distance as in (1).

4. Experimental Results

For all experiments two sets of speaker independent multiple mixture monophone HMMs have been used, one trained on British English as the target language and one set trained on Latin-American Spanish representing the source language. The models were build with the HTK Toolkit ([7]) and the regression class trees were built using the techniques described in [4].

In Figure 1 the weights for a global transformation of both one native and two Spanish accented speakers are compared. The

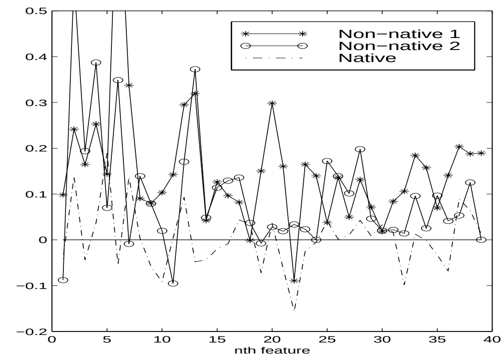

Figure 1: Comparison of model combination weights for non-native and native speakers

weights for the native speaker are distributed around 0.0 and generally are smaller in magnitude than the weights of the non-native speakers. This indicates that an improved HMM for foreign accented speech might be found through the combination of models from the source and the target language. Likewise, the model combination approach might have little impact on improving the modelling of a native speaker, since the weight will not change much of the original models.

In the following experiments recognition accuracy has been measured using a system with a word-pair grammar. This grammar was based on the stories in simplified English used for the recording of a non-native database. This data was taken from a specifically recorded database of non-native speech of students of English as a foreign language, [6]. Each recognition test contained 90 sentences per speaker.

In Figure 2 the recognition performance for baseline, MLLR and model combination with all three mapping types is shown as a function of the number of adaptation sentences. Both the fact that the performance does not increase with more adaptation sentences and that adapted models can perform worse than the baseline indicate how different the non-native data are to the native target. Because recognition results did not increase significantly with increased adaptation data, the number of adaptation sentences has been limited to 5. Furthermore, this represents a reasonable number of adaptation sentences to ask for from a user of a CALL system.

Recognition performance for MLLR and the model combination technique with all three different mappings can be seen in Table 1. For all speakers the model combination technique yields a maximum improvement of 5% over the baseline performance and of 3.4% for MLLR. Comparing the different mapping techniques it can be seen that the model-level mapping, which is based on the error knowledge of a student, yields the best results. For the state- and mixture level mapping the performance is similar to MLLR. These results indicate that incorporating knowledge about typical pronunciation mistakes can help to improve non-native recognition. For the bilingual approach the results are better than MLLR and two of the mappings, but worse than the model-based mapping.

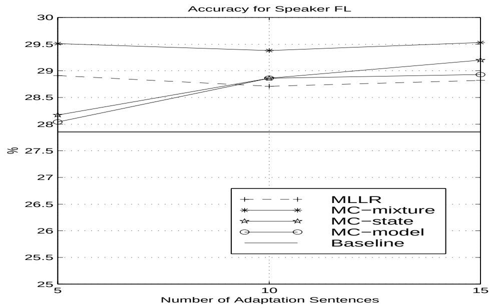

Figure 2: Phone Recognition results for MLLR and Model Combination as a function of adaptation sentences

Spkr	Base	MLLR	MC1	MC2	MC3	MC-Bi
FL	65.93	66.33	70.54	66.53	67.40	70.01
PC	57.12	58.87	60.06	58.37	59.44	58.93
TS	64.10	64.49	65.50	65.18	64.04	65.69
aver.	62.38	63.23	65.37	63.29	63.63	64.45

Table 1: Recognition Accuracy for different mapping approaches (MC1: Model-level mapping, MC2:State-level, MC3:Mixture-level, MC-Bi:Bilingual Alignment (with regression tree)), 5 adaptation sentences

The last experiment uses a-priori combined models. When choosing the a-priori weights, the following aspects were taken into account. Firstly, the combination weights for non-native speakers tend to be in the interval $[0, 0.2]$, thus it was decided to use weights $b_i = 0.1$. Secondly, since foreign accent especially causes changes in the second and higher formats of accented speech, the first 6 weights, which represent the six lowest mel-frequency cepstral coefficients, have been set to zero. The results are shown in Table 2. The choice of weights used in this setup represent an educated guess. It will need further experiments to determined optimal values. However, with this approach the results for MLLR can be improved without any additional computational loads, since the combination can be done off-line.

5. Conclusions

A technique for combining speaker independent models of the target and source language for non-native accents has been presented. In these preliminary experiments, the basic technique of model combination yield slight recognition improvement over the standard MLLR adaptation technique. Also, some improvement has been obtained by using a-priori combined models. In this case the same recognition performance as with MLLR could be obtained by using off-line combined models. This means MLLR equivalent can be achieved without adaptation data if the type of accent is known.

Speaker	Baseline	A-priori	MLLR	A-priori MLLR
FL	65.93	66.33	66.07	70.74
PC	57.12	58.43	59.31	58.37
TS	64.10	67.74	65.57	64.68
aver.	62.38	63.17	63.65	64.40

Table 2: Recognition accuracy for baseline, a-priori baseline, MLLR, a-priori based MLLR and a-priori model combination (model-level), 5 adaptation sentences

Further experiments will be needed to explore this type of accent adaptation more thoroughly. Future work will use cross-word triphone models instead of monophone models in order to improve the alignment accuracy. Finally, investigations are necessary on the integration of the adaptation technique presented here in computer-assisted language learning systems.

6. Acknowledgements

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

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, E. Knodt, Khudanpur S., 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 *Speech Technology in Language Learning (STiLL*, pages 37–40. ESCA Workshop, May 1998.
3. L.L. Chase. *Error-Responsive Feedback Mechanisms for Speech Recognizers*. PhD thesis, Carnegie Mellon University, Pittsburgh, USA, 1997.
4. M.J.F. Gales. The generation and use of regression class trees for mllr adaptation. Cued/f-infeng/tr263, Cambridge University Engineering Department, 1996.
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 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.
7. S.J. Young, J. Odell, D. Ollason, and P.C. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory, 1996.