SPEECH RECOGNITION OF MULTIPLE ACCENTED ENGLISH DATA USING ACOUSTIC MODEL INTERPOLATION

Thiago Fraga-Silva[*†], *Jean-Luc Gauvain*[*], *Lori Lamel*[*]

[*] LIMSI - CNRS, Spoken Language Processing Group, Orsay, France
[†] Université Paris-Sud, Orsay, France
{thfraga, gauvain, lamel} @limsi.fr

ABSTRACT

In a previous work [1], we have shown that model interpolation can be applied for acoustic model adaptation for a specific show. Compared to other approaches, this method has the advantage to be highly flexible, allowing rapid adaptation by simply reassigning the interpolation coefficients. In this work this approach is used for a multi-accented English broadcast news data recognition, which can be considered an arduous task due to the impact of accent variability on the recognition performance. The work described in [1] is extended in two ways. First, in order to reduce the parameters of the interpolated model, a theoretically motivated EM-like mixture reduction algorithm is proposed. Second, beyond supervised adaptation, model interpolation is used as an unsupervised adaptation framework, where the interpolation coefficients are estimated on-the-fly for each test segment.

Index Terms— Model interpolation, supervised and unsupervised adaptation, multi-accented data.

1. INTRODUCTION

It is well-known that speech variability is an important difficulty in automatic speech recognition, directly affecting the quality and challenging the robustness of the recognition systems. Many sources of variability exist, such as gender, accent, age, speaking style and rate of speech [2]. According to [3], accent is, behind gender, the second principal source of speech variability. Even if substantial progress has been made in the techniques for normalization and speaker adaptation to compensate some of the variability, recognition accuracy has been shown to strongly degrade when the accent of the test speaker is not well represented in the training data [4–6].

A general approach to deal with accented data is to adapt a prior model to the target accent [5]. This approach can be extended to the multi-accented data case [6,7]. The main principle is to build different accent-dependent models and use a selector for adaptation. This procedure was adopted by [8] to recognize multiple accented English data. In that work, accent-specific models were created for 6 different geographical regions where English is spoken as an official language. Similar to [9], a Gaussian mixture model (GMM) based classifier was used to select an accent-dependent model for each test segment. On average, a significant improvement was obtained over the accent-independent system, although the accuracy of one of the accents (Middle-East [1]) degraded.

This paper revisits the problem addressed in [8]. However, instead of using the approach based on maximum *a posteriori* (MAP) adaptation [10], here we propose to build the accent-dependent models by interpolation, with coefficients automatically estimated [1] via an expectation-maximization (EM) algorithm [11]. Model interpolation has been already applied for non-native speech recognition tasks [5, 12], but with manually selected coefficients.

Model interpolation is performed by merging the GMMs of the component models and properly adjusting the mixture coefficients. As a consequence, the number of parameters of the model increases according to the number of component models used, which also increases decoding time. To recover the same computational complexity, a GMM reduction algorithm is required. As described in [13], the mixture reduction issue is usually defined as an optimization problem where the objective function is an arbitrary similarity measure between the original and the reduced models. However, we propose here a different theoretical point of view that fits into the EM framework and leads to a *soft-clustering* algorithm. In that sense, this algorithm has something in common with the (hard-) clustering algorithm proposed by [14]. Like the latter, we also initialize the reduced mixtures with a greedy algorithm [15]. Nevertheless, we do not perform the last optimization step, since it degraded the recognition performances.

In this work, interpolation was assessed in two ways. First, the accent-dependent models are built via interpolation during the training phase. Then, during decoding, a model selection is applied. Second, the component models are interpolated on-the-fly, with coefficients estimated for each test segment. Therefore, instead of making a hard decision for an accent, a smoothed combination is performed.

This work has been partially supported by OSEO, the French State agency for innovation, under the Quaero program.

[1] In [8], the labels for the ME and NA accents were mistakenly swapped.

The remainder of this paper is organized as follows. Section 2 describes the interpolated model, while Section 3 presents the proposed GMM reduction algorithm. In Section 4, the corpus and the baseline system are described. Section 5 describes the experiments performed with the interpolated models. Conclusions are given in Section 6.

2. GAUSSIAN MIXTURE MODEL INTERPOLATION

Linear interpolation is perhaps one of the most straightforward manner to combine models. In this work, the GMMs of different component acoustic models are interpolated. To do so, it is considered that the component models have the same structure and have been independently estimated.

Let us denote the GMM parameters of the component models by $\{\lambda_k^{(l)}\}_{k=1\ldots K}^{l=1\ldots L}$, where the indices refer to the l-th state of the k-th component model. Thus, it is assumed that we dispose of K component models with L states each. The likelihood of a sample x_t, observed at time t, given the interpolated model of state s can be represented by:

$$f(x_t|\lambda^{(s)}) = \sum_{k=1}^{K} \alpha_k^{(s)} \cdot f(x_t|\lambda_k^{(s)}) \tag{1}$$

where $\{\alpha_k^{(s)}\}_{k=1\ldots K}$ are the interpolation coefficients, with $\sum_k \alpha_k^{(s)} = 1$. The models $\lambda_k^{(s)}$ are GMMs with probability density functions expressed by:

$$f(x_t|\lambda_k^{(s)}) = \sum_{i=1}^{M_k^{(s)}} \omega_{ki}^{(s)} \cdot \mathcal{N}\left(x_t \middle| \mu_{ki}^{(s)}, \Sigma_{ki}^{(s)}\right) \tag{2}$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ is a normal density function with mean vector μ and covariance matrix Σ, and M is the number of mixture components in the model. By inspection of (1) and (2), it follows that the interpolated model $\lambda^{(s)}$ is also a GMM, with $\sum_k M_k^{(s)}$ components and mixture weights $\left\{\alpha_k^{(s)}\omega_{ki}^{(s)}\right\}_{k=\{1\ldots K\}, i=\left\{1\ldots M_k^{(s)}\right\}}$.

The main advantage of such a method is that the model can be easily adapted to a new task, show or speaker through simple reassignment of the interpolation coefficients. Moreover, these coefficients can be manually set or automatically estimated via an EM algorithm, which maximizes the likelihood on some held-out data [1]. Alternatively, the coefficients can be estimated during decoding, from the test data itself. In this latter case, the model interpolation can be seen as an unsupervised adaptation method.

The method has something in common with the cluster adaptive training [16] and eigenvoices [17], which have been demonstrated to belong to the same family of adaptive training methods [18]. These methods attempt to model the acoustic variation with respect to the available training subsets, by performing a joint parameter estimation. Thus, any speaker is constrained to be represented by a combination of these subsets. However, in the case of model interpolation, the models are estimated independently. Furthermore, any existent model judged to be relevant for a task can be considered for interpolation. This aspect makes interpolation a more flexible approach to be used on adaptation. Moreover, as performed in this work, model interpolation can be used on top of speaker adaptive training [19], also an adaptive training method.

3. GAUSSIAN MIXTURE REDUCTION

As mentioned before, the number of parameters in the interpolated model increases proportionally to the number of used component models, what may also affect the decoding time. In order to reduce system complexity, a GMM reduction algorithm is required. The algorithm proposed here is theoretically motivated from the maximum likelihood estimation (MLE) on the training data.

Let us consider two multivariate GMMs, one with M components and parameters $\bar{\lambda}$, and its reduced version with N components ($N < M$) and parameters λ:

$$\bar{\lambda} = \left(\bar{\omega}_1 \ldots \bar{\omega}_M, \bar{\mu}_1 \ldots \bar{\mu}_M, \bar{\Sigma}_1 \ldots \bar{\Sigma}_M\right) \tag{3}$$

$$\lambda = \left(\omega_1 \ldots \omega_N, \mu_1 \ldots \mu_N, \Sigma_1 \ldots \Sigma_N\right) \tag{4}$$

where $\bar{\omega}_i$, $\bar{\mu}_i$ and $\bar{\Sigma}_i$ denote, respectively, the mixture coefficient, the mean vector and the covariance matrix for the i-th Gaussian component of the model $\bar{\lambda}$, and ω_j, μ_j and Σ_j denote the same parameters for the j-th component of λ.

Let us consider the MLE of the model λ, but with the additional constraint that each observation vector x_t is restricted to a cluster $\bar{\phi}_i$. In this manner, the vectors do not contribute independently to the estimation of the reduced model, but indirectly via the clusters themselves. Doing so, it can be shown that the so-called auxiliary function becomes:

$$Q(\lambda, \hat{\lambda}) = \sum_{i=1}^{M}\sum_{j=1}^{N} \bar{n}_i \cdot \gamma_{ij} \cdot \log\left(\hat{\omega}_j \cdot f\left(\bar{\phi}_i|\hat{\mu}_j, \hat{\Sigma}_j\right)\right) \tag{5}$$

where $\bar{n}_i$ is the number of vectors associated to the cluster $\bar{\phi}_i$ and

$$\gamma_{ij} = f\left(j|\bar{\lambda}_i, \lambda\right) = \frac{\omega_j \cdot f\left(\bar{\phi}_i|\mu_j, \Sigma_j\right)}{\sum_{j'} \omega_{j'} \cdot f\left(\bar{\phi}_i|\mu_{j'}, \Sigma_{j'}\right)} \tag{6}$$

Now, considering that the cluster $\bar{\phi}_i$ has a normal distribution defined by the parameters of the original model, that is $\bar{\phi}_i \sim \mathcal{N}(\cdot|\bar{\mu}_i, \bar{\Sigma}_i)$, we can define the likelihood $f\left(\bar{\phi}_i|\mu_j, \Sigma_j\right)$ using the following conditional expectation:

$$\begin{aligned}\log f\left(\bar{\phi}_i|\mu_j, \Sigma_j\right) &:= E\left[\log f\left(x|\mu_j, \Sigma_j\right)|\bar{\phi}_i\right] \\ &= \log \mathcal{N}(\bar{\mu}_i|\mu_j, \Sigma_j) - \frac{1}{2}\mathrm{tr}\left({\Sigma_j}^{-1}\bar{\Sigma}_i\right)\end{aligned} \tag{7}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

Set	US	AU	GB	ME	NA	IN
training (hours)	316	33	55.4	27.7	8.2	9.4
train6h (hours)	6	6	6	6	6	6
held-out (hours)	3.6	2.0	2.0	1.6	2.2	1.7
test (min)	172	19	45	15	15	15

Table 1. Duration information of the multi-accented English corpus used here. *train6h* and *held-out* are subsets of *training*. Accents: United States (US), Australia (AU), Great Britain (GB), Middle East (ME), North Africa (NA), India (IN).

Normalizing the auxiliary function by the number of training samples, we can substitute $\bar{n}_i$ by $\bar{\omega}_i$ in (5) without any other changes. Finally, the parameter update equations can be obtained by partially deriving the auxiliary function with respect to the model parameters and setting them to zero:

$$\hat{\omega}_j = \sum_{i=1}^{M} \bar{\omega}_i \gamma_{ij} \tag{8}$$

$$\hat{\mu}_j = \frac{1}{\hat{\omega}_j} \sum_{i=1}^{M} \bar{\omega}_i \gamma_{ij} \cdot \bar{\mu}_i \tag{9}$$

$$\hat{\Sigma}_j = \frac{1}{\hat{\omega}_j} \sum_{i=1}^{M} \omega_i \gamma_{ij} \cdot \left((\bar{\mu}_i - \hat{\mu}_j)(\bar{\mu}_i - \hat{\mu}_j)^\top + \bar{\Sigma}_i \right) \tag{10}$$

These equations are similar to those used in [14], except for the term γ_{ij}, which allows each cluster $\bar{\phi}_i$ to be associated to different Gaussians λ_j with a certain probability. In this sense, this algorithm can be seen as a *soft-clustering* method, and, by construction, as a constrained or smoothed MLE.

Different methods for initialization have been tried (single Gaussian, highest coefficient components, random), but better results have been achieved by initializing the reduced mixtures with the Runnall's algorithm [15].

4. DATA AND SYSTEM OVERVIEW

The data and baseline system used in this work are the same as presented in [8]. They are briefly described in this section.

4.1. Corpus

Table 1 shows the duration of training and test sets of the broadcast news English corpus used in this work, containing 6 different accents. The true accent label for each speaker is unknown. Here, the geographical region from where the show was broadcast is considered as the accent label for all the speakers in the audio file. The *train6h* and *held-out* sets are non-overlapping randomly selected subsets of the training data. They were used to estimate the interpolation coefficients for each accent. The size of the test subsets was selected based on the distribution of data available for training. The audio comes from a variety of news sources, mostly collected via satellite with some downloaded from the Web.

4.2. System overview

The broadcast speech recognition system used in this work is quite similar to other systems developed at LIMSI [20]. It uses a 42 dimensional PLP-likeacoustic feature vector, including 12 cepstrum coefficients, log energy and pitch, along with their first and second derivatives.

The phone set contains 35 phones, as well as special units for silence, breath and hesitation markers. Silence is modeled by a GMM with 1024 components. The other units are modeled by context-dependent triphone hidden Markov models, where each state observation is modeled by a GMM with 32 components. The models cover about 18k phone contexts and contain 11.5k tied states. They are gender dependent, speaker adapted [19], and were obtained via maximum mutual information estimation (MMIE) [21].

The language models (LMs) were trained on about 1.2 billion words of texts coming from various sources, including news and transcriptions. They use a 65k-word vocabulary list and were built by interpolation of LM components estimated on different subsets of the training data.

Decoding is performed using unsupervised maximum likelihood linear regression (MLLR) and constrained MLLR for speaker adaptation [22]. A word lattice is generated in a first step, followed by a LM rescoring procedure and a consensus decoding [23].

The baseline models are described in [8] and have been generated as follows. First, an accent-independent gender-independent model was created. Gender-specific and accent-specific models were obtained using a joint MAP adaptation, followed by one iteration of MMIE. During decoding, the most likely accent-dependent model is selected based on a GMM classifier for each test segment. For the case where the decision has been performed at the show level, a 100% classification accuracy rate has been obtained. The main results obtained by [8] are represented in the first part of Table 2. On average, the multi-accent system (with per show accent-ID) performs better than the accent-independent system. However, the performance for the ME accent deteriorates.

5. EXPERIMENTS

In this work, we restrain our analysis to acoustic modeling, assuming that the remaining system components (pronunciation dictionary, language models) provide a reasonable coverage of all the accents represented in the data set. However, it is known and expected that better recognition performances can be obtained by adapting the overall system.

For the remaining experiments, the accent-dependent models built for the baseline system were used as the component models for interpolation.

Method	US	AU	GB	ME	NA	IN	Sum	Ave
Accent independent	14.34	11.92	12.84	15.90	26.47	39.28	16.07	20.12
Show-accent-ID	13.95	11.91	11.98	16.46	25.19	34.28	15.39	18.96
Interpolated	13.83	11.55	11.08	15.79	24.29	33.52	**15.05**	**18.34**
Interpolated + reduced	13.75	11.87	11.45	15.79	24.89	33.95	15.11	18.62
Speaker-interpolation	14.11	11.37	11.65	15.63	24.24	33.18	15.27	18.36
Show-interpolation	14.06	11.18	11.30	15.86	24.24	33.69	15.22	18.39

Table 2. WER(%) results using different recognition systems for each of the 6 regional accents. 'Sum' corresponds to the WER on the whole test set, while 'Ave' to the average WER when each subset is weighted equally. The first part shows results for the baseline models; the second and third parts, for models interpolated, either during the training (2nd), or decoding (3rd) phase.

5.1. Supervised adaptation via model interpolation

Model interpolation was assessed in a supervised adaptation scheme. First, the *train6h* subset was used to estimate context-independent models for each accent. Interpolation coefficients were estimated in order to maximize the likelihood on the respective *held-out* data. With the estimated coefficients, the component models were interpolated, generating accent-specific models. These interpolated models had their GMMs reduced to 32 components. During decoding, a GMM classifier was used to select the accent-specific interpolated model for each show. The results obtained with this approach are shown in the second part of Table 2.

On average, the interpolated models led to the best word recognition performances on the test data. With respect to the two baseline systems (Show-accent-ID and Accent-independent), relative improvements of 2% and 6%, respectively, have been achieved. The use of the GMM reduction algorithm led to a slight loss of performance. However, in both cases, the interpolated models led to better recognition performances for all of the 6 test accents. In particular, and contrary to the Show-accent-ID system, the performance on the ME data do not degrade when compared to the accent-independent system. The improvements obtained with the interpolated models are significant with respect to both baseline systems, according to the NIST matched pair sentence-segment word error significance test [24].

5.2. Unsupervised adaptation via model interpolation

To assess model interpolation as an unsupervised adaptation technique, the following procedure was performed. During decoding, instead of selecting one accent, interpolation coefficients are estimated on-the-fly by maximizing the likelihood of the test data itself. With these coefficients, a segment-specific model is created by interpolation of the component models. In this case, no GMM reduction is applied. Speaker or show-specific interpolated models were created. The results are shown in the third part of Table 2.

At a first glance, this approach leads to similar average performances compared to the Show-accent-ID system. However, the performance is worse for the US accent and better for all the other accents. This may be explained by the fact that the US accent is better represented in the training data. Thus, the unsupervised decision leads to a loss of performance. With this prior knowledge, it would be possible to build a hybrid system, that performs hard (selection) or soft (interpolation) decisions based on the accent representativeness on the training data. Alternatively, only some of the component models could be chosen for interpolation, based on a threshold decision applied to the interpolation coefficients.

System	test	train6h (as a test)
Accent-independent	18.77	17.22
Adapted to *held-out*	19.15	17.39
Interpolated	17.25	15.86
Interpolated + reduced	17.45	16.33
Speaker-interpolation	**17.09**	**15.71**

Table 3. WER(%) results on the ME test and train6h sets for the case where no ME data has been considered available for acoustic model training. No MMIE was performed.

5.3. Leaving target accent out

Another set of experiments was performed to assess the case where the target accent data is not represented in the training corpus. Deliberately, the ME accent was chosen. For this accent, it was assumed that only the *held-out* set was available. The same procedure as before was used to create the component models, but without considering the ME training set. Moreover, MMIE was not used here, since it led to worse recognition performances. The systems were evaluated on the ME *test* and *train6h* sets. Since this latter set was not used on training, its inclusion on evaluation was feasible.

Table 3 summarizes the results obtained. In general, the same conclusions drawn for the ME data from the previous experiments can be confirmed here, but with more significant gains. For instance, in the case where interpolation was applied as a supervised adaptation technique (3rd row), relative improvements of about 8%–9% were obtained with respect to the accent-independent system. With the reduced interpolated model, the relative gains are around 5%–7%. Finally, slightly better results were obtained using the unsupervised adaptation scheme (last row).

6. CONCLUSION

In this work, acoustic model interpolation was assessed for a multiple-accented English data recognition task. Compared to a previously proposed approach [8] that uses MAP adaptation, it led to a small, but significant, gain of performance for all the 6 accents represented in the corpus. As an extension to a previous work [1], this paper has also presented a theoretically correct solution to reduce the number of Gaussian components of the interpolated models.

A major advantage of the interpolation method is its flexibility in the sense that it can be used to rapidly adapt a system to a new target by simply reassigning the interpolation coefficients. This characteristic also allows the use of interpolation as an unsupervised adaptation technique, which leads to competitive performance levels compared to the baseline system. This technique is especially interesting when the target accent is not represented in the training data.

REFERENCES

[1] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Interpolation of acoustic models for speech recognition," in *Interspeech*, Lyon, France, 2013, pp. 3347–3351.

[2] M. Benzeghiba et al., "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[3] C. Huang, T. Chen, S. Z. Li, E. Chang, and J.-L. Zhou, "Analysis of speaker variability," in *Eurospeech*, 2001, pp. 1377–1380.

[4] V. Fischer, E. Janke, S. Kunzmann, and T. Ross, "Multilingual acoustic models for the recognition of non-native speech," in *Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 331–334.

[5] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *ICASSP*, 2003, vol. 1, pp. I–540.

[6] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *Interspeech*, 2005, pp. 217–220.

[7] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2-3, pp. 141–153, 2004.

[8] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic speech recognition of multiple accented English data," in *Interspeech*, Makuhari, Japan, 2010, pp. 1652–1655.

[9] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 343–346.

[10] J.-L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[12] T. P. Tan and L. Besacier, "Acoustic model interpolation for non-native speech recognition," in *ICASSP*, 2007, vol. 4, pp. IV–1009.

[13] D. F. Crouse, P. Willett, K. Pattipati, and L. Svensson, "A look at Gaussian mixture reduction algorithms," in *International Conference on Information Fusion*, 2011.

[14] D. Schieferdecker and M. F. Huber, "Gaussian mixture reduction via clustering," in *International Conference on Information Fusion*, 2009, pp. 1536–1543.

[15] A. R. Runnalls, "Kullback-Leibler approach to Gaussian mixture reduction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, 2007.

[16] M. Gales, "Cluster adaptive training for speech recognition," in *ICSLP*, 1998, vol. 1998, pp. 1783–1786.

[17] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation.," in *ICSLP*, 1998, vol. 98, pp. 1774–1777.

[18] M. Gales, "Multiple-cluster adaptive training schemes," in *ICASSP*, 2001, vol. 1, pp. 361–364.

[19] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*, 1996, vol. 2, pp. 1137–1140.

[20] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[21] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *ICASSP*, 1986, vol. 11, pp. 49–52.

[22] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, pp. 171, 1995.

[23] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization," in *Eurospeech*, 1999.

[24] D. S. Pallet, W. M. Fisher, and J. G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *ICASSP*, 1990, pp. 97–100.