FEATURE NORMALIZATION USING SMOOTHED MIXTURE TRANSFORMATIONS

Patrick Kenny, Vishwa Gupta, Gilles Boulianne, Pierre Ouellet and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

`{Patrick.Kenny, Vishwa.Gupta, Gilles.Boulianne, Pierre.Ouellet, Pierre.Dumouchel}@crim.ca`

ABSTRACT

We propose a method for estimating the parameters of SPLICE-like transformations from individual utterances so that this type of transformation can be used to normalize acoustic feature vectors for speech recognition on an utterance-by-utterance basis in a similar manner to cepstral mean normalization. We report results on an in-house French language multi-speaker database collected while deploying an automatic closed-captioning system for live broadcast news. An unusual feature of this database is that there are very large amounts of training data for the individual speakers (typically several hours) so that it is very difficult to improve on multi-speaker modeling by using standard methods of speaker adaptation. We found that the proposed method of feature normalization is capable of achieving a 6% relative improvement over cepstral mean normalization on this task.

Index Terms: speech recognition, SPLICE, eigenvoice MAP

1. INTRODUCTION

If feature normalization is ever to compensate for all sources of speaker, environmental and channel variability encountered in speech recognition then it has to be applied on an utterance-by-utterance basis. Our aim in this paper is to show how eigenvoice methods, which have previously been applied only in the model domain, can be used to derive non-linear transformations for feature normalization in speech recognition that can be applied to individual utterances in much the same as cepstral mean normalization.

The transformations that we study are smoothed mixture transformation of which one of the earliest examples is SPLICE (originally developed to deal with environmental mismatches in the Aurora databases [1]). The novelty in this paper consists in showing that the parameters of such transformations can be robustly estimated using much smaller amounts of data than have previously been envisaged. We report results on an in-house French language multi-speaker database collected while deploying an automatic closed-captioning system for live broadcast news. An unusual feature of this database is that there are very large amounts of training data for the individual speakers (typically several hours) so that it is very difficult to improve on multi-speaker modeling by using standard methods of speaker adaptation. We found that the proposed method of feature normalization is capable of achieving a 6% relative improvement over cepstral mean normalization on this task.

2. MIXTURE TRANSFORMATIONS

A wide variety of acoustic feature normalization techniques (including codeword-dependent cepstral normalization, vector Taylor series methods and the feature mappings used to compensate for channel effects in speaker recognition) can be viewed as instances of a type of transformation which is defined by a Gaussian mixture model (GMM) and a collection of offset vectors, one for each Gaussian in the GMM. If, for each mixture component c in the GMM, the corresponding offset vector is denoted by o_c, the type of transformation we are considering is given by $Y \to X$ where

$$X = Y - o_c \tag{1}$$

and c is the mixture component that accounts for Y.

We will refer to such a transformation as a *mixture transformation*. Note that, in this formulation, the role of the GMM is to model the distribution of un-normalized speech (Y) rather than normalized speech (X). In the original SPLICE algorithm, multiple environment-specific GMM's were used to model unnormalized speech but, since our goal is to carry out feature normalization on an utterance-by-utterance basis, we will use a single GMM for this purpose. This GMM can be thought of as a universal background model (UBM) in the sense in which this term is used in speaker recognition.

We will use the term *smoothed mixture transformation* to refer to a variation on this theme which has received widespread attention recently. The idea is to define a a transformation $Y \to X$ where

$$X = Y - \sum_c P(c|Y)o_c \tag{2}$$

and, for each mixture component c, $P(c|Y)$ denotes the posterior probability of the event that Y is generated by sampling from the Gaussian distribution corresponding to this component. (These are the probabilities that are calculated in Baum-Welch training of GMM's. If (1) is viewed as a stochastic transformation then the right hand side of (2) is the expectation of the right hand side of (1) conditioned on Y.) This is a smooth non-linear transformation which is very similar to multi-dimensional interpolation using radial basis functions.

Smoothed mixture transformations have been used to compensate for environmental noise [1, 2] and inter-speaker variation [3] and to enhance the phonetic discrimination capability of speech recognizers [4, 5]. In these applications, relatively large amounts of data are available to estimate the offset vectors o_c and research has focused on finding the best criteria for this estimation problem. (Maximum likelihood, maximum mutual information, minimum classification error and minimum phone error have all been studied.)

3. THE CONNECTION WITH CEPSTRAL MEAN NORMALIZATION

In this paper we are concerned with another type of application which can be motivated by considering the case where there is just one Gaussian in the UBM so that the transformation (2) is defined by a single offset vector. If, for a given utterance, the offset vector is estimated by averaging the acoustic features over the length of the utterance then applying the smoothed mixture transformation (2) to the observations in the utterance is just cepstral mean normalization (CMN). In the general case, (2) can be viewed as a type of cepstral subtraction in which the vector to be subtracted varies from one observation to another (since the posterior probabilities $P(c|Y)$ in (2) depend on Y). This raises the question of whether (2) can be applied on an utterance-by-utterance basis in the same way as ordinary CMN.

Implementing this requires estimating a set of offset vectors o_c for an arbitrary utterance. We propose to estimate the offset vectors in such a way that applying (1) or (2) to the observations in the given utterance will cause them to be distributed according to the UBM, just as applying CMN to an arbitrary utterance causes the observations in the utterance to be distributed with mean 0. This is a reasonable objective because, if the UBM is trained with a sufficiently large and diverse database, then speaker, environmental and channel effects will be averaged out and the mean vectors in the UBM will reflect only phonetic variability. So, ideally, normalized utterances will be stripped of speaker, environmental and channel effects. Histogram normalization techniques such as Gaussianization are also based on the idea of mapping feature vectors onto canonical distributions. In these approaches the individual features are treated as if they were statistically independent and the mapping is implemented by means of scalar quantization. The idea underlying our approach is to use (soft) vector quantization instead.

Of course, just as in CMN, care is needed in handling very short utterances because it is not possible to estimate the offset vectors reliably in this situation. Similarly, in the case of a UBM having a large number of Gaussians, so that a large number of offset vectors have to be estimated, the estimation problem is not straightforward unless the 'utterances' are extraordinarily long. Indeed the problem is equivalent to estimating utterance dependent GMM's which have the same number of Gaussians as the UBM. For if m denotes the supervector obtained by concatenating the Gaussians in the UBM and M the supervector derived from an utterance dependent GMM then

$$M = m + o \qquad (3)$$

where o is the supervector obtained by concatenating the offset vectors o_c. Since m is known, the problem of estimating o is equivalent to the problem of estimating M. So the estimation problem can be thought of either as a problem of estimating a set of utterance dependent offset vectors or as a problem of estimating an utterance dependent GMM and we will use these two perspectives interchangeably.

Maximum likelihood estimation is not capable of dealing with this type of estimation problem if the number of Gaussians is large. The results in [3] suggest that, even with 5 minutes of data per speaker, there is nothing to be gained by using a UBM with more than 64 Gaussians if maximum likelihood is used as the estimation criterion. Similarly, the authors in [2] limited themselves to a UBM with only 32 Gaussians. This is consistent with the evidence from text-independent speaker recognition where MAP estimation has supplanted maximum likelihood estimation so that GMM's with larger numbers of Gaussians can be handled.

There are some applications of (2) in which extremely large numbers of Gaussians have proved to be useful but there does not seem to be any possibility of implementing the corresponding estimation algorithms on an utterance-by-utterance basis [4, 5]. In the core test of the annual NIST text-independent speaker recognition evaluations, the enrollment data for a target speaker consists of single utterance (such as a Switchboard conversation side). The most widely used approach to speaker modeling is to use a UBM with 1–2 K Gaussians and MAP estimation to estimate a speaker-dependent GMM of the same size from each target speaker's enrollment utterance. This is the approach that we will take here.

Two flavors of MAP estimation that have been used in speaker recognition are classical MAP [6, 7] and eigenvoice MAP [8]. The strengths and weaknesses of these types of MAP estimation complement each other (see [9] and the references cited there for a framework which embraces both). Classical MAP estimation requires large amounts of training data and it is asymptotically equivalent to maximum likelihood estimation. Because very large numbers of eigenvoices cannot be robustly estimated in practice, there is no such guarantee for eigenvoice MAP but it is generally recognized that a modest number of eigenvoices does give good estimates with small amounts of training data. Thus eigenvoice MAP is a natural choice for the estimation problem confronting us.[1]

Basing the procedure for estimating the offset vectors in (2) on eigenvoice MAP gives explicit control over the number of free parameters that have to be estimated. (The number of free parameters involved in estimating the eigenvoices is RCF, where R is the number of eigenvoices, C the number of mixture components in the UBM and F the dimension of the acoustic feature vectors. These parameters are estimated from the same database as the UBM. In estimating the offset vectors for a given utterance, the number of free parameters is just R.) Moreover, the estimation procedure is guaranteed to be well behaved even in the case of very short utterances (this is the primary reason for using a Bayesian rather than a maximum likelihood estimation criterion). In both of these respects, our approach to feature normalization differs from FMLLR (also known as constrained MLLR), which can only be used for utterance-by-utterance normalization in situations where the utterances to be recognized are sufficiently long that they can be used to reliably estimate affine transformations of the feature space by maximum likelihood methods. It also differs from standard FMLLR in that it does not make use of phonetic transcriptions so that it can be applied blindly at recognition time. (We say 'standard FMLLR' because the need for transcriptions or word graphs can be avoided by implementing FMLLR with a GMM rather than a HMM as proposed in [10].)

We will use the term *smoothed mixture normalization* (SMN) to refer to utterance-by-utterance feature normalization based on (2) where the offset vectors for each utterance are estimated by eigenvoice MAP.

---

[1] We are guilty of an abuse of language here: eigenvoice MAP as it is presented in [8] is designed to produce speaker dependent models whereas we are confronted with the task of estimating *utterance* dependent GMM's. So eigenvoice MAP has to be implemented with utterances playing the role of speakers. This distinction is only pertinent in situations where multiple utterances have been recorded for a speaker.
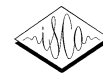
4. EXPERIMENTS

4.1. Training and test sets

Our experiments were conducted on a multi-speaker database comprising 22 speakers which was collected while deploying an automatic closed-captioning system for live broadcast news in French. (This system is described in a companion paper [11].) The training data for our experiments consisted of 39 hours of data collected between October 2004 and April 2005. The test set consisted of 16 hours of data collected from 20 of the 22 speakers between May and October 2005. This is a very large test set (94 K words) so that small differences in recognition accuracies may be statistically significant. The training and test utterances for this task are of highly variable duration (mostly the range 30 seconds – 5 minutes). Utterances of duration 1 minute or more account for most of the data and we restricted ourselves to these utterances in designing our training and test sets.

Our experience with this test set has been that inter-speaker variation is very well modeled by using mixture distributions having reasonably large numbers of mixture components (e.g. 64) so that we have been unable to obtain performance improvements by using standard speaker adaptation techniques such as MLLR and MAP. For example, in an experiment involving four female speakers the only improvement we were able to obtain was an increase in the percentage of words correctly recognized from 88.2 to 88.3.

4.2. Implementation of CMN

For signal processing we used 12 cepstral coefficients and a log energy feature together with their first and second derivatives calculated every 10 ms. Silences were removed using a slightly modified version of the public domain ISIP silence detector.

We implemented a causal version of CMN in order to minimize the delay in producing real-time recognition decisions. The simplest strategy is to estimate a mean vector μ_0 for each speaker by averaging over the speaker's training data. A better approach is to estimate an utterance dependent mean vector for each utterance, by taking μ_0 as the initial estimate and updating it as successive frames become available. This leads to the following normalization procedure which we refer to as *real-time* CMN and which we implemented with $\alpha = 0.005$:

$$\mu_t = (1-\alpha)\mu_{t-1} + \alpha Y_t$$
$$X_t = Y_t - \mu_t \tag{4}$$

Here, Y_t represents the unnormalized feature vector at time t and X_t the normalized feature vector. Explicitly,

$$\boldsymbol{\mu}_t = (1-\alpha)^t \mu_0 + \alpha \sum_{\tau=1}^{t} (1-\alpha)^{t-\tau} Y_\tau \tag{5}$$

so that μ_t is a weighted average of μ_0 and the observations up to time t, with the contribution of μ_0 decaying exponentially over time.

Real-time CMN can be applied to the cepstral coefficients $c_1, \ldots, c_{12}$ alone (as in the HTK implementation of file-based CMN) or it can be applied to the energy feature as well. Our experience has been that the latter approach is much more effective. In the the special case where $\alpha = 0$, the mean vector μ_0 is not updated so we will refer to this case as *non-adaptive* CMN.

4.3. Implementation of SMN

Since SMN (like FMLLR) is basically an off-line procedure we did not attempt to address the causality problem. Except where otherwise indicated, we applied CMN on a file-by-file basis (HTK style) and we used the 39-dimensional features obtained in this way as the starting point for SMN.

We used 200 eigenvoices and a UBM having 512 Gaussians (estimated using the multi-speaker training set). As explained in Section 3, we formulate the problem of estimating the offset vectors for a given utterance as one of estimating an utterance dependent GMM supervector M and use the methods in [8]. This entails first estimating a rectangular matrix v of low rank so that the prior distribution of M has mean m and covariance matrix vv^*. (The columns of v can be interpreted as eigenvoices.) Then for each utterance, the posterior distribution of M can be calculated using the observations in the utterance.

The calculations are described in Propositions 1 and 2 of [8] which can be applied directly provided that the following condition is satisfied: For each observation vector Y and each mixture component c,

$$P(c|X) = P(c|Y) \tag{6}$$

where X and Y are related as in (1). If this condition is not satisfied then this problem can be dealt with by the iterative procedures described in Section IV of [8]. We have observed modest performance improvements by doing this type of iteration so we used it in all of our experiments with SMN.

4.4. Results

For each experiment we rebuilt a decision tree having roughly 1000 leaf nodes and we used 64 Gaussians per mixture distribution throughout.

Our multi-speaker results are summarized in Table 1. Comparing lines 1, 2 and 3 shows that real-time CMN is very effective and a good gain in performance can be obtained by applying it to the energy feature as well as to the the cepstral coefficients. The best result was obtained with SMN (line 4) which represents a 6% relative improvement over real-time CMN (line 3). The result in line 5 shows that SMN works better with file-based CMN than with real-time CMN which is not surprising since SMN is an offline method. Table 2 breaks out the result in line 4 over the 20 test speakers. The results here show that there are a few goats among our speaker set who have resisted our attempts at feature normalization.

Table 1. *Different types of feature normalization. Multi-speaker results. The test set consists of 94 K words uttered by 20 speakers. Homophone confusions are counted as errors.*

	Adaptation Type	Accuracy (%)
1	Non-adaptive CMN including energy	78.1
2	Real-time CMN excluding energy	80.9
3	Real-time CMN including energy	82.5
4	**File-based CMN + SMN**	**83.6**
5	Real-time CMN including energy + SMN	83.0

Table 2. *Breakdown of the result in line 4 of Table 1 showing the number of words spoken by each speaker (N), the percentage of words correctly recognized and the percentage accuracies. Homophone confusions are counted as errors.*

speaker	Correct (%)	Accuracy (%)	N
abr	84.09	79.89	2859
ala	92.33	90.12	7370
cla	87.64	85.47	833
dle	87.83	85.06	5670
dup	83.84	80.97	9462
imi	87.48	85.56	727
jbe	88.51	86.14	7763
jfb	85.35	82.47	6596
jub	88.99	86.92	4495
kba	90.34	88.38	6345
kbl	90.74	88.50	3174
kbr	81.03	77.35	6146
lec	86.30	83.37	7701
lha	87.88	82.81	1617
mfr	92.46	90.15	5715
mnj	80.71	76.64	762
mru	82.70	77.35	8006
sgo	68.72	64.61	7129
sla	90.64	89.18	1368

5. DISCUSSION

In this paper we have shown how to use eigenvoice methods to estimate smoothed mixture transformations for feature normalization in speech recognition that can be applied to individual utterances in the same way as cepstral mean normalization. We obtained a 6% (relative) reduction in error rates on a multi-speaker task using this method. On the face of it, this is not a very large improvement but since there are no perceptible mismatches between the training and test set in our experimental set up and standard methods of speaker adaptation have proved to be ineffective on this task, this is a satisfying result. It would be interesting to experiment with smoothed mixture normalization on more challenging tasks where there is substantial speaker and channel variability such as the Switchboard and Fisher databases.

Eigenvoice methods were originally developed for HMM model adaptation in speech recognition but we have had a good deal more success in applying them to GMM's in text-independent speaker recognition [9]. The reason for this appears to be that when eigenvoice methods are applied with HMM's rather than GMM's, errors in phonetic transcriptions and weaknesses in decision tree based triphone modeling introduce a source of variability which is not low-dimensional (contrary to the primary assumption of eigenvoice modeling). This motivated us to study the question of whether it might be possible to find a way of using eigenvoice based estimation of GMM's in speech recognition by working in the feature domain rather than the model domain. The advantages of working in the feature domain are well known, namely that there is no need to create speaker or utterance adapted HMM's at recognition time and speaker adaptive training can be carried out very easily. The approach that we have developed has the additional advantages that it can be applied blindly at recognition time (since it does not rely on phonetic transcriptions) and it behaves robustly on short utterances (since it is based on MAP estimation).

The problem we have addressed in this paper is that of feature normalization for speech recognition where the goal is to remove both speaker and channel effects from individual utterances. Feature normalization for speaker recognition presents a rather more subtle challenge since the goal in this situation is to remove channel effects but keep speaker effects intact. In a very interesting independent development, the authors in [12] have shown how methods similar to those presented here can be successfully applied to this problem.

6. REFERENCES

[1] L. Deng, J. Wu, J. Droppo, and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms," *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 477–480, June 2005. [Online]. Available: http://research.microsoft.com/srg/papers/2005-deng-spl.pdf

[2] J. Wu, Q. Huo, and D. Zhu, "An environment compensated maximum likelihood training approach based on stochastic vector mapping," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.

[3] K. Visweswariah and P. Olsen, "Feature adaptation using projection of Gaussian posteriors," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005.

[4] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FMPE: discriminatively trained features for speech recognition," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.

[5] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005.

[6] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[8] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification," in *Proc. ICASSP 2006*, Toulouse, France, May 2006.

[10] G. Stemmer, F. Brugnara, and D. Giulani, "Adaptive training using simple target models," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005.

[11] G. Boulianne, P. Cardinal, F. Osterrath, P. Ouellet, *et al.*, "Computer-assisted closed-captioning of live TV broadcasts in French," in *Proc. ICSLP*, Pittsburgh, Pennsylvania, Sept. 2006.

[12] C. Vair, D. Coibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. IEEE Odyssey 2006*, San Juan, Puerto Rico, June 2006.