MAXIMUM-LIKELIHOOD AFFINE CEPSTRAL FILTERING (MLACF) TECHNIQUE FOR SPEAKER NORMALIZATION

*Yoon Kim**

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, Stanford, CA 94305 USA
yoonie@ccrma.stanford.edu

ABSTRACT

We present a novel technique of minimizing the acoustic variability of speakers by transforming the features extracted from the speaker's data to better fit the recognition model. The concept of maximum-likelihood affine cepstral filtering (MLACF) will be introduced for feature transformation, along with solutions for the transformation parameters that maximize the likelihood of the test data with respect to a given recognition model. It is shown that for log-concave distributions, the solution of the MLACF problem can be obtained using convex programming. HMM-based digit recognition on the TIDIGITS database is presented to demonstrate the flexibility of the transformation in compensating for large acoustic mismatches between the speakers in the training and test database. In addition, it will be shown that the technique requires estimation of far fewer transformation parameters compared to existing techniques, thus allowing fast, real-time compensation.

1. INTRODUCTION AND MOTIVATION

One of the challenges in speaker-independent recognition is normalizing the acoustic variability across speakers due to differences in physiology (e.g., pitch, vocal-tract geometry) as well as other factors. The problem of compensating for differences in individual speaker characteristics can be attacked using two different approaches – *normalization* and *adaptation*. Speaker normalization usually refers to a process of transforming the feature space so that the acoustic mismatch between the model and the data is minimized. Speaker adaptation tries to alter the parameters of the speech recognition model to fit the data in hand.

The ideas behind speaker normalization are mainly based on signal processing concepts that stem from the physical model of the vocal tract. It is well accepted that the vocal tract can be modeled as a set of cascaded acoustic tubes, and each tube in turn can be viewed as a resonator, having a distinct set of resonant frequencies that depend on the shape and size of the tube. Recognizing that the frequencies of the vocal-tract resonance — formants — are crucial to the perception of phones, researchers have developed methods to compensate for differences in vocal-tract length. The technique is referred to as vocal tract length normalization (VTLN), and it has been the predominant choice for speaker normalization. VTLN is typically performed by warping the linear frequency axis in an attempt to align the formant frequencies for a better spectral match between data from different speakers [1, 2, 3].

Speaker adaptation has been viewed as a model transformation technique to maximize the stochastic matching between the model and the data. Most of the speaker adaptation methods use some optimality criterion to estimate the transformation of a model Λ_X to a new model Λ_Y. One simple method is to restrict the adaptation to be a fixed bias in the cepstral domain [4]. The most widely used technique for speaker adaptation however, is *maximum-likelihood linear regression* (MLLR) [5, 6]. It uses a set of regression-based transforms represented by an $L \times L$ matrix (L is the feature dimension) to tune the HMM mean parameters to a new speaker. The estimation of the MLLR matrix parameters is done using the Baum-Welch procedure.

When the amount of adaptation data is large enough for reliable estimation of all L^2 coefficients of the matrix, MLLR serves as a very powerful tool for modeling complex transformations in the cepstral space. However, if adaptation data is limited, the number of parameters to be estimated has to be reduced by imposing certain constraints on the transformation matrix (e.g., diagonal, block-diagonal, triangular) [6, 7]. Unfortunately, these constraints are not based on signal processing intuition.

In the following, a cepstral normalization technique is presented for deriving a speaker-dependent transformation that maximizes the likelihood of a given feature set with respect to a pre-trained stochastic model. The transformation corresponds to multiplicative filtering plus bias in the log spectral domain. The filtering has an effect of modifying the spectral properties, in particular the spectral tilt and the overall shape of the spectral envelope. We note that in addition to frequency warping for VTLN, cepstral filtering further normalizes speaker acoustics by altering the spectral tilt and overall shape. These are considered to be important attributes of speaker characteristics along with formant locations [8, 4], and spectral tilt is usually affected by glottal characteristics which depend on the individual speaker acoustics. Figure 1 shows the effects of cepstral filtering in the log spectral domain when applied to vowel normalization.

The parameters of the transformation correspond to a speaker-dependent compensation filter and bias that contains information specific to each speaker. It will be shown that due to the Toeplitz structure of the transformation matrix, estimating the maximum likelihood (ML) parameters for normalization can be posed as a convex optimization problem for a special class of probability distributions, including the multivariate Gaussian distribution. It will also be shown that for the proposed technique, the number of parameters to be estimated is $\mathcal{O}(L)$ rather than $\mathcal{O}(L^2)$ as in MLLR. Moreover, for models with Gaussian distributions, the parameters can be estimated using a closed-form formula, with only a fraction of the computation required for MLLR estimation.

* The author is now with VerbalTek Inc., Santa Clara, CA USA

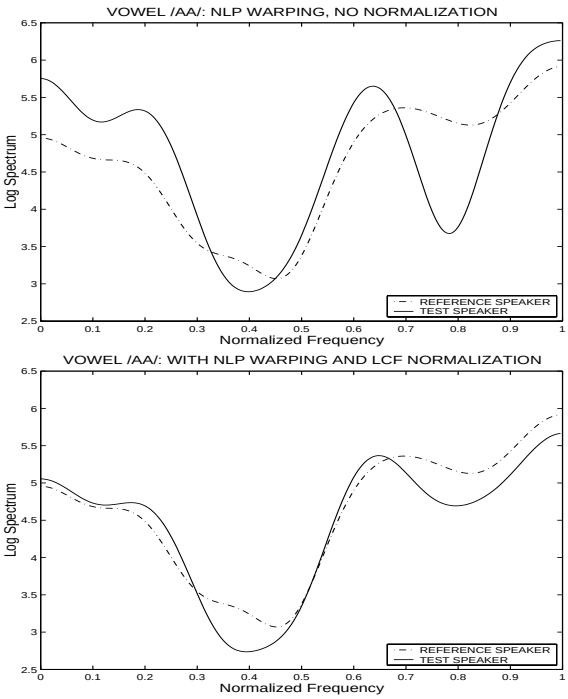

Fig. 1. Effects of cepstral filtering normalization in the frequency domain for the vowel /AA/. The top figure represents the mel-smoothed vowel spectrum of a test speaker (solid) along with the reference mel spectrum for the same vowel (dash-dotted). The bottom figure shows the two spectra after cepstral filtering is applied to the mel spectrum of the test speaker.

2. ML AFFINE CEPSTRAL FILTERING

2.1. Problem Formulation

Let $\mathcal{C} = \{c^{(0)}, c^{(1)}, \ldots, c^{(N-1)}\}$ be the sequence of cepstral features obtained from the speech data of a certain speaker, where N is the total number of features for the dataset, and $c^{(i)} \in \mathbf{R}^L$. Also assume that each feature vector belongs to a generalized phone class, where the number of classes is K, and that each phone class is associated with a model $\Lambda_k, k = 1, \ldots, K$. Note that a generalized phone can be an extension of a phone (e.g., diphone, triphone, phonetically tied clusters). The affine transformation that maps $c^{(i)}$ to $\tilde{c}^{(i)}$ is defined as follows:

$$\tilde{c} = Hc + v, \tag{1}$$

$$\begin{aligned} H &= \text{conv}(h) = \text{conv}(h_0, h_1, \ldots, h_{L-1}) \\ &\triangleq \begin{bmatrix} h_0 & 0 & 0 & \cdots & 0 \\ h_1 & h_0 & 0 & \cdots & 0 \\ h_2 & h_1 & h_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ h_{L-1} & h_{L-2} & h_{L-3} & \cdots & h_0 \end{bmatrix} \end{aligned}$$

The above affine transformation is nothing but a Toeplitz matrix representation of convolution pertaining to a causal, linear, time-invariant filtering plus a bias:

$$\tilde{c}_n = \sum_{k=0}^{L-1} c_k \, h_{n-k} + v_n, \quad n = 0, 1, \ldots L-1. \tag{2}$$

In the frequency domain, we have

$$\tilde{S}(e^{j\omega}) = H(e^{j\omega})S(e^{j\omega}) + V(e^{j\omega}), \tag{3}$$

where $\tilde{S}(e^{j\omega})$ and $S(e^{j\omega})$ are the log spectra pertaining to $\tilde{c}$ and c respectively, and $H(e^{j\omega})$ and $V(e^{j\omega})$ are the Fourier transform of the filter h and bias v respectively. Affine cepstral filtering can thus be viewed as spectral shaping in the log domain, with $H(e^{j\omega})$ being the speaker-specific, shaping filter and $V(e^{j\omega})$ being the additive bias. We note that due to the commutativity of convolution, we can also express the transformed features in terms of the filter and bias as

$$\tilde{c} = Hc + v = Ch + v = \hat{C}x, \tag{4}$$

$$C = \text{conv}(c), \quad \hat{C} = [C \quad I], \quad x = [h \quad v]^T. \tag{5}$$

The transformed feature $\tilde{c}$ can thus be expressed as a linear function of x.

2.2. MLACF Solution

Based on the above framework, our objective is to find the parameter $x^* = [h^* \quad v^*]^T$ that optimally maps the features of a speaker to match the model being used for recognition. Let's consider a stochastic model associated with each feature. For each feature vector, there exists a corresponding output probability distribution $P(c; \Lambda_c)$, where c is the observation (feature) vector, and Λ_c denotes the corresponding model for c. The transformation of features from $\mathcal{C}$ to $\tilde{\mathcal{C}} = \{\tilde{c}^{(0)}, \tilde{c}^{(1)}, \ldots, \tilde{c}^{(N-1)}\}$ will result in a change in the observation probability, and in turn in the overall likelihood. Define the *total log-likelihood* of $\tilde{\mathcal{C}}$ as [1]

$$\mathcal{L}_{\tilde{\mathcal{C}}}(x) \triangleq \sum_{i=0}^{N-1} \mathcal{L}_{\tilde{c}}^{(i)}(x), \tag{6}$$

where

$$\begin{aligned} \mathcal{L}_{\tilde{c}}^{(i)}(x) &= \log P(\tilde{c}^{(i)}; \Lambda_{k_i}) \\ &= \log P(Hc^{(i)} + v; \Lambda_{k_i}), \\ &= \log P(\hat{C}^{(i)}x; \Lambda_{k_i}), \quad i = 0, \ldots, N-1. \end{aligned} \tag{7}$$

Here, k_i is the phone-class index which the feature vector $c^{(i)}$ belongs to, and $\hat{C}^{(i)} = [C^{(i)} \quad I]$ as in Equation 5.

The objective is to find the MLACF parameter x that results in the transformed features having maximum total-likelihood. Since

[1] When a random variable is transformed, the relationship between the log-likelihood of the transformed variable $\tilde{c}$ and the original variable c can be expressed as $\log g(\tilde{c}) = \log f(c) - \log J$, where $g(\cdot)$ and $f(\cdot)$ are the PDFs of $\tilde{c}$ and c respectively, and J is the Jacobian of the transformation from c to $\tilde{c}$. However, the basic assumption of feature normalization in this paper is that the PDF corresponding to the test feature is *different* from that of the training data (acoustic mismatch), and the transformed test feature $\tilde{c}$ is assumed to have the PDF corresponding to the trained model that we already have. Thus, no Jacobian term was added in the computation of the transformed likelihood in Equation 7.

the logarithm function is a monotonically increasing function, the maximum-likelihood objective can be also written in terms of log-likelihood as

$$\max_x [\mathcal{L}_{\tilde{c}}(x)] \Longleftrightarrow \min_x [-\mathcal{L}_{\tilde{c}}(x)]. \tag{8}$$

If the negative log-likelihood is a *convex* function in x, the solution is unique and can be obtained efficiently using convex optimization techniques. In summary, for any probability distribution that is *log-concave* in $\tilde{c}$, the negative log-likelihood function $-\mathcal{L}_{\tilde{c}}(x)$ is *convex*, since $\tilde{c}$ is a linear function of x and linearity preserves convexity. Fortunately, many useful probability distributions are log-concave. One important example is the multivariate Gaussian distribution, which is discussed in the next section.

2.3. Case: Gaussian Distribution

Let's assume the PDF is a single Gaussian of the form

$$P(c; \Lambda_k) = \frac{1}{((2\pi)^L |\Sigma_k|)^{1/2}} e^{-\frac{1}{2}(c-\mu_k)^T \Sigma_k{}^{-1}(c-\mu_k)}, \tag{9}$$

where Λ_k denotes the model for class k, and μ_k and Σ_k are the mean and covariance of Λ_k, respectively. The total log-likelihood of the transformed features becomes

$$\mathcal{L}_{\tilde{c}}(x) = \sum_{i=0}^{N-1} \mathcal{L}_{\tilde{c}}^{(i)}(x), \tag{10}$$

where

$$\begin{aligned} \mathcal{L}_{\tilde{c}}^{(i)}(x) &= \log P(\tilde{c}^{(i)}; \Lambda_{k_i}) = \log P(\hat{C}^{(i)}x; \Lambda_{k_i}) \\ &= \kappa_i - \frac{1}{2}(\hat{C}^{(i)}x - \mu_{k_i})^T \Sigma_{k_i}^{-1} (\hat{C}^{(i)}x - \mu_{k_i}). \end{aligned}$$

It is easily seen that $\mathcal{L}_{\tilde{c}}^{(i)}(x)$ is concave in x. The objective function to be minimized is a sum of quadratic forms in x which depend on the data and the model, plus constants κ_i. The problem of finding the MLACF coefficients x for a given speaker can be posed as the following convex optimization problem:

$$\text{minimize} \quad f(x) = \sum_{i=0}^{N-1} (\hat{C}^{(i)}x - \mu_{k_i})^T \Sigma_{k_i}^{-1} (\hat{C}^{(i)}x - \mu_{k_i}), \tag{11}$$

where $\hat{C}^{(i)}$, μ_{k_i} and Σ_{k_i} are the cepstral matrix, mean and covariance for the model corresponding to $c^{(i)}$ respectively.

The summation in Equation 11 can be expressed as a sum of squared ℓ_2-norms, as

$$f(x) = \sum_{i=0}^{N-1} \|\Sigma_{k_i}^{-1/2}(\hat{C}^{(i)}x - \mu_{k_i})\|_2{}^2. \tag{12}$$

The objective function can be rewritten as

$$f(h) = \|Ax - b\|_2{}^2, \tag{13}$$

$$A = \begin{bmatrix} \Sigma_{k_0}^{-1/2} C^{(0)} & \Sigma_{k_0}^{-1/2} \\ \Sigma_{k_1}^{-1/2} C^{(1)} & \Sigma_{k_1}^{-1/2} \\ \vdots & \vdots \\ \Sigma_{k_{N-1}}^{-1/2} C^{(N-1)} & \Sigma_{k_{N-1}}^{-1/2} \end{bmatrix} \in \mathbf{R}^{NL \times 2L},$$

$$b = \begin{bmatrix} \Sigma_{k_0}^{-1/2} \mu_{k_0} \\ \Sigma_{k_1}^{-1/2} \mu_{k_1} \\ \vdots \\ \Sigma_{k_{N-1}}^{-1/2} \mu_{k_{N-1}} \end{bmatrix} \in \mathbf{R}^{NL}.$$

It is easily seen that the solution is given by

$$x^* = \begin{bmatrix} h^* \\ v^* \end{bmatrix} = (A^T A)^{-1} A^T b. \tag{14}$$

For Gaussian mixture densities

$$P(c; \Lambda) = \sum_{i=1}^{M} w_i N(c; \mu_i, \Sigma_i), \quad w_i \geq 0, \quad \sum_{i=1}^{M} w_i = 1,$$

we use the approximation

$$P(c; \Lambda) \approx \max_i [w_i N(c; \mu_i, \Sigma_i)]. \tag{15}$$

The above approximation has been widely used in HMM-based recognition to increase speed and efficiency. Also, some researchers reported the merit of using Gaussian mixtures with a small number of components for performing vocal-tract normalization [1].

3. EXPERIMENTAL RESULTS

We conducted a HMM-based digit recognition experiment using the TIDIGITS speech corpus, with 326 speakers (111 men, 114 women, 50 boys and 50 girls) providing approximately 77 digit sequences per speaker.

3.1. Case I: Adult Data on Adult Model

First, using the training set of 112 adult speakers (55 men, 57 women), baseline adult models were created. An 8-state HMM for each digit was trained using 1–15 Gaussians per state. In other words, we varied the number of Gaussians per state, and created 15 models per each digit for experimental purposes. The features used were the NLP cepstrum [9] and MFCC. Then, using the test data from 113 speakers (56 men, 57 women), we performed recognition using the baseline model. After the baseline recognition, we used one utterance per speaker per digit to estimate the normalization parameters. Using the estimates obtained for each speaker, MLACF transformation was performed on the rest of the data, and the transformed features were used in recognition.

Table 1 shows the word error rate (WER) results for HMM digit recognition of adult data using adult models, for the cases of unnormalized (baseline) and MLACF. The results obtained when using separate models for male and female data (gender-dependent models) are also shown for comparison. Note that the results shown are the *best* results obtained by varying the number of Gaussians from 1 to 15. Values in parentheses represent the minimum number of Gaussians used to obtain the best-case results.

We see that MLACF normalization achieves a 26 percent reduction of WER compared to the baseline case for both the NLP and the MFCC features. Note that while the MFCC and NLP techniques produce similar results for the unnormalized case, normalized NLP features yielded slightly lower error rates than those of MFCC features. Also, the normalized models required less Gaussians per mixture.

Table 1. WER for digit recognition of adult data using adult models for the cases of unnormalized (baseline), MLACF, and when using gender-dependent (GD) models. Values in parentheses represent the minimum number of Gaussians per mixture used to obtain the best-case results.

Feature	Baseline	MLACF	GD
MFCC	1.9 (4)	1.5 (4)	1.8
NLP	1.8 (15)	1.4 (9)	1.8
Avg	1.9	1.4	1.8

Table 2. WER for digit recognition of child data using adult models for the cases of unnormalized (baseline), MLACF, and when using a child model to test child data. Values in parentheses represent the minimum number of Gaussians per mixture used to obtain the best-case results.

Feature	Baseline	MLACF	Child
MFCC	20.0 (2)	15.5 (5)	4.5
NLP	19.5 (2)	15.6 (5)	2.3
Average	20.0	15.5	3.4

3.2. Case II: Child Data on Adult Model

Now we consider a case of more severe mismatch between the training and test data. A dataset of 50 children (25 boys, 25 girls) was used as the test data using the adult models obtained in the previous experiment. Table 2 shows the results for the cases of unnormalized, and MLACF normalization, along with those using a model trained with child data. Again the results shown are the best results obtained by varying the number of Gaussians from 1 to 15. Comparing the results with those obtained in Table 1, it is evident that significant acoustic mismatch between child and adult data resulted in a catastrophic increase in error rate by a factor of 10. The MLACF technique resulted in an error rate reduction of 23 percent relative, even in the case of severely mismatched speaker acoustics. However, there is still a large room for improvement, when compared with the child model case. Note that for the child-adult case, the optimal number of Gaussians for the baseline case was 2, implying that in severely mismatched conditions, merely adding Gaussians is not enough to improve performance.

4. CONCLUSION

We introduced a new method of compensating speaker acoustics for robust speech recognition. Speaker normalization was described as a process of minimizing the acoustic mismatches between a given set of data and a recognition model, by transforming the features to exhibit maximum likelihood with respect to a given acoustic model. A special affine transformation in the cepstral space was proposed for transforming the features derived from the speech segment. By modeling the transformation as filtering in the log spectrum domain plus a bias, the normalization was shown to be an affine cepstral transformation. Furthermore, the normalization matrix was shown to be triangular and Toeplitz (namely a convolution matrix), reducing the number of parameters to be estimated from $\mathcal{O}(L^2)$ to $\mathcal{O}(L)$ where L is the dimension of the feature. For log-concave distributions (e.g., Gaussian PDFs), the problem of estimating the optimal normalization parameters that maximize the likelihood of the data with respect to an existing model was shown to be convex. Applying MLACF normalization gave large wins in recognition performance over the baseline system in HMM digit recognition using the TIDIGITS database (26 percent WER reduction). The results of digit recognition where child data was tested on an adult model (23 percent WER reduction) showed that MLACF is indeed an effective technique even in cases of large mismatch of speaker acoustics.

5. REFERENCES

[1] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. ICASSP*, 1999, vol. 2, pp. 761–764.

[2] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, January 1998.

[3] J. W. McDonough, W. Byrne, and X. Luo, "Speaker normalization with all-pass transforms," in *Proc. of ICSLP*, 1998.

[4] S. Cox and J. Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting," in *Proc. ICASSP*, 1989, pp. 294–297.

[5] 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, pp. 171–185, 1995.

[6] V. Digalakis, D. Rtischev, and L. Neumeyer, "Fast speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, pp. 357–366, September 1995.

[7] E. Bocchieri, V. Digalakis, A. Corduneanu, and C. Boulis, "Correlation modeling of MLLR transform biases for rapid HMM adaptation to new speakers," in *Proc. ICASSP*, 1999, vol. 2, pp. 773–776.

[8] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. of Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[9] Y. Kim and J. O. Smith, "A speech feature based on bark frequency warping – The non-uniform linear prediction (NLP) cepstrum," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.