Upper Bound Kullback-Leibler Divergence for Hidden Markov Models with Application as Discrimination Measure for Speech Recognition

Jorge Silva and Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory, http://sail.usc.edu
Department of Electrical Engineering, Viterbi School of Engineering
University of Southern California
jorgesil@usc.edu, shri@sipi.usc.edu

Abstract— **This paper presents a criterion for defining an upper bound Kullback-Leibler divergence (UB-KLD) for Gaussian mixtures models (GMMs). An information theoretic interpretation of this indicator and an algorithm for calculating it based on similarity alignment between mixture components of the models are proposed. This bound is used to characterize an upper bound closed-form expression for the Kullback-Leibler divergence (KLD) for left-to-right transient hidden Markov models (HMMs), where experiments based on real speech data show that this indicator precisely follows the discrimination tendency of the actual KLD.**

I. Introduction

The Kullback-Leibler divergence (KLD) and its symmetric extension, the divergence, provide objective statistical indicators for the difficulty in discriminating between two statistical hypotheses [1]. In its original formulation these hypotheses are characterized by two probability density functions (pdfs), f_1 and f_2, where the Kullback-Leibler distance between hypotheses H_1 with respect to H_2 is given by:

$$D\left(\mu_1 || \mu_2\right) = \int f_1(o) \log \frac{f_1(o)}{f_2(o)} \partial o \qquad (1)$$

where $\mu_2 \gg \mu_1$ is a necessary condition for the KLD to be well defined [1][1]. In this context, $D\left(\mu_1 || \mu_2\right)$ is the average information per observation to discriminate H_1 with respect to H_2 [1]. The KLD can be used to compare probabilistic models from a discrimination point of view, and to globally evaluate the inherent discrimination complexity in a recognition task [2]. Those are some of the reasons that explain their wide use in the context of classification based on decision-theoretic approaches [3], [4].

In spite of its clear definition and meaningful interpretation, even for the case of hypotheses modeled as probability density functions (pdfs), the KLD has closed-form expression but for a very limited family of models, such as multivariate Gaussian and the generalized Gaussian distributions [3], [1]. In particular, there is no closed-form KLD expression for the important family of Gaussian mixture models (GMMs). Vasconcelos [4] has recently addressed the problem of approximating the KLD for the case of GMMs. This work presents closed-form expression for the case of vector quantizer (VQ) density estimators and proposes an approximation for GMMs based on asymptotic likelihood approximation (ALA) property, which is proved to be close to the actual KLD under some strong intra-discrimination condition between the mixture components. In addition, an upper bound was proposed by Singer et al [5] based on the log-sum inequality [6].

The problem is even more intricate if we consider more complex probabilistic structures like hidden Markov models (HMMs). These models can be considered a generalization of GMMs in infinite dimensional observation space and consequently the KLD does not have a closed-form expression [2]. For this problem, Singer et al [7] propose an upper bound for KLD for the case of discrete observation HMMs and Do [8], an upper bound for the KLD rate (KLDR) for continuous ergodic HMMs. We recently extended those upper bound characterizations for the important case of left-to-right *transient* continuous density HMMs [9], which are the models used to represent basic sub-word units in automatic speech recognition (ASR) [10]. In this context, this current work can be considered an extension of [9] that provides an improved characterization of the closed-form upper bound at the observation level of the models, which is the model fraction that captures most of the discrimination information between HMMs [11].

This paper presents a connection between the ALA approximation framework [4] and the upper bound for GMMs presented in [5]. An alignment criterion between the mixture components of the models is derived to characterize the proposed upper bound. In addition, this paper presents an information theoretic interpretation of the upper bound based on the characterization of the unknown joint mixture-observation space distribution. We use this upper bound KLD between GMMs to characterize a closed-form upper bound KLD between left-to-right *transient* HMMs. The representation quality of this indicator is evaluated by correlating it with a numerical approximation of the KLD and some applications are presented within a automatic speech recognition scenario.

[1] μ_i is the probability measure absolutely continuos with respect to the Lebesgue measure in $\mathbb{R}$ and f_i its Radon-Nikodym derivative.

II. Upper Bound KLD for left-to-right Transient HMMs

Hidden Markov sources can be characterized as two statistically dependent family of random sequences [12]. The first family is the observable random process $\{O_n\}_{n\in\mathbb{Z}_+}$, which represents observations or measurements of a physical phenomenon. For the context of this work, a single random variable of this process O_n takes values in a continuous finite dimensional space $\mathcal{O} = \mathbb{R}^N$. The second family is the underlying state process $\{X_n\}_{n\in\mathbb{Z}_+}$ which is a first order homogeneous Markov source with individual random variables taking values in a finite-alphabet or state space denoted by $\mathcal{X} = \{1, 2, .., S\}$ [13]. This Markov process is not observable but is statistically dependent with $\{O_0, O_1, ...\}$, and it is used to indirectly model the dynamic evolution of the marginal probability distribution of the observable process. The statistical dependency between these two random sequences is given by the following relationship:

$$P(O_{j_1} = o_1, .., O_{j_L} = o_L | X_{j_1} = x_1, .., X_{j_L} = x_L) = \prod_{k=1}^{L} P(O_{j_k} = o_k | X_{j_k} = x_k) \quad (2)$$

$\forall L > 0$, $\forall (j_1, j_2, .., j_L) \in (\mathbb{Z}_+)^L$, $\forall (x_1, x_2, .., x_L) \in \mathcal{X}^L$, $\forall (o_1, o_2, .., o_L) \in \mathcal{O}^L$. This relationship represents the fact that, conditional to an observation of the random variable X_t, the observable random variable associated with the same time index O_t is independent of any other random variable of the hidden Markov source. Given this property, the distribution of the observable process $\{O_0, O_1, ...\}$ can be completely characterized based on the Markov distribution of the underlying state process and the family of conditional observation distribution.

The representation of the observable source distribution, $P_{O^{\mathbb{Z}_+}}(\cdot)$, is called a hidden Markov model (HMM) and it is denoted by $\lambda = (\pi, A, B)$, where $\pi \equiv (P(X_0 = x))_{x\in\mathcal{X}}$ represents the initial state probability distribution, $A \equiv (P(X_{k+1} = x_2 | X_k = x_1))_{(x_1,x_2)\in\mathcal{X}^2}$ is the time-invariant stochastic matrix of the underlying Markov chain and $B \equiv \left(b_x(\cdot) \equiv P_{O_k/X_k=x}(\cdot)\right)_{x\in\mathcal{X}}$ is the family of time-invariant conditional observation distributions, [10].

Let us consider two distributions for the joint process $\{(O_n, X_n)\}_{n\in\mathbb{Z}_+}$ with the following standard parametric representation $\lambda^1 = \left(\pi^1, A^1, B^1\right)$ and $\lambda^2 = \left(\pi^2, A^2, B^2\right)$, [10]. In this scenario where all states have associated a observation distribution, [9] shows the KLD between λ_1 and λ_2 is upper bound by the following expression:

$$D\left(\lambda^1 || \lambda^2\right) \equiv \lim_{n\to\infty} H_{\lambda^1||\lambda^2}\left(O_0^{n-1}\right) \leq \lim_{n\to\infty} D\left(\pi^1 || \pi^2\right) + \pi^{1^t}\left(\sum_{i=0}^{n-2} A^{1^i}\right)\left(d^A_{KLD} + d^B_{KLD}\right) + \pi^{1^t} A^{1^{n-1}} d^B_{KLD} \quad (3)$$

where $d^A_{KLD} \equiv \left(D\left(A^1_x || A^2_x\right)\right)_{x\in\mathcal{X}}$ is the column vector whose entries are the KLD between the row probability mass function of the stochastic matrices of λ_1 and λ_2, and $d^B_{KLD} \equiv \left(D\left(b^1_x || b^2_x\right)\right)_{x\in\mathcal{X}}$ is the column vector with the KLD between the the observation distribution of the models.

A. Left-to-right transient HMMs

An extension of the above mentioned source model, used to represent random process with finite number of observations, is to consider hidden Markov processes with an additional family of states which do not generate observations [14], [9]. Those states are called non-emitting. Under this context, the original state space $\mathcal{X}$ can be partitioned into a family of emitting states $\mathcal{X}_o$ and non-emitting states $\mathcal{X}_u$.

A particular case of those models is the family of left-to-right HMMs with final non-emitting and absorbing state, the standard models used to represent sub-word phoneme units in speech recognition [10], [9] and the practical focus of this exposition. Left-to-right HMMs satisfy the afore mentioned *transient* property. The reason is that the only recurrent state is non-emitting and under this condition left-to-right HMMs generate finite number of observations almost surely, independent of their initial state distribution.

For this family of left-to-right transient HMMs, [9] proves that if $\pi^2 \gg \pi^1$; $A^2_j \gg A^1_j$, $\forall j \in \mathcal{X}_o$ and $b^2_x \gg b^1_x$, $\forall x \in \mathcal{X}_o$ then the KLD is well defined and has an upper bound given by:

$$D\left(\lambda^1 || \lambda^2\right) \leq D\left(\pi^1 || \pi^2\right) + \hat{\pi}^{1\ t} \cdot (I - \hat{A}^1)^{-1} \cdot \left(\hat{d}^A_{KLD} + \hat{d}^B_{KLD}\right) \quad (4)$$

where $\hat{d}^B_{KLD} = \left(D\left(b^1_x || b^2_x\right)\right)_{x\in\mathcal{X}_o}$ and $\hat{d}^A_{KLD} = \left(D\left(A^1_x || A^2_x\right)\right)_{x\in\mathcal{X}_o}$ are column vectors and $\hat{A}^1$ is the stochastic matrix of the model λ^1 restricted to the emitting states.

Note that in order to have a closed-form upper bound from Eqn.(4), we need a closed-form expression for the KLD between the observation distribution of the HMMs involved. Considering the standard continuous observation HMMs used in speech recognition scenarios [10], [14], this implies finding a closed-form upper bound for the KLD between Gaussian mixture models (GMMs). The next section addresses this point motivated by the approach presented in [4].

III. KLD Approximation for GMMs

Let us consider two GMMs $f^l(x) = \sum_{k=1}^{K} w^l_k \cdot \mathcal{N}(x, \mu_{l,k}, \Sigma_{l,k})$, $l \in \{1, 2\}$ defined in the same finite dimensional space $\mathcal{O}$. Vasconcelos [4] proposes the asymptotic likelihood approximation (ALA) given by:

$$D_{ALA}\left(f_1 || f_2\right) \equiv \sum_{k=1}^{K} w^1_k \log \frac{w^1_k}{w^2_{\beta(k)}} + \sum_k w^1_k \cdot D\left(\mathcal{N}(x, \mu_{1,k}, \Sigma_{1,k}) || \mathcal{N}(x, \mu_{2,\beta(k)}, \Sigma_{2,\beta(k)})\right) \quad (5)$$

where $(\beta(k))_{k\in\{1,..,K\}}$ is an alignment based on the similarity between the mixture components, motivated by the asymptotic scenario when mixtures of the same model have almost perfect discrimination and, under some additional conditions, this approximation turns out to be arbitrarily close to actual the KLD [4]. However, the mentioned asymptotic intra-discrimination assumption is unrealistic for Gaussian mixtures models, because Gaussian distributions have infinite support and the general practical application of those models does not guarantee this intra-discrimination assumption. Consequently, it is not possible to characterize any formal theoretical relationship between the ALA approximation and the actual KLD in the general case. Motivated by that, we propose to constrain this alignment approach, $\beta(\cdot)$ to be in addition an injective mapping, where by the log-sum inequality [6], Eqn.(5) results to be an upper bound of the KLD between general mixture models [5], [8], [9]. In particular, we characterize $\beta(\cdot)$ by:

$$\beta(k) = \arg \qquad (6)$$
$$\min_{\hat{k}\in\{1,..,K\}\setminus I(k)} D\left(\mathcal{N}(x,\mu_{1,k},\Sigma_{1,k})||\mathcal{N}(x,\mu_{2,\hat{k}},\Sigma_{2,\hat{k}})\right)$$

where $I(k) = \{\beta(j) : j < k\}$, $1 < k < K$, and $I(1) = \phi$, that ensures an injective ALA similarity alignment. Note that for the case of Gaussian models, the KLD has a closed-form expression [1] and hence this alignment criterion can be efficiently implemented. The next section provides an alternative information theoretic interpretation of the afore mentioned upper bound KLD for GMMs.

A. Revisiting the Upper Bound for Mixture Models

Let us consider a random vector O defined in $\mathbb{R}^n$ and a discrete random variable X defined in a finite alphabet $\mathcal{X} = \{1, .., K\}$. Suppose that M and P are two probability measures for the joint measurable space $(\mathbb{R}^n \times \mathcal{X})$. Then, the relative entropy of the joint random vector (O, X) with measure P with respect to M is given by [1] (*Corollary 3.3 pp. 18*):

$$\begin{aligned} H_{P||M}(O, X) &= \sum_{k\in\mathcal{X}} P(X = k) \cdot \log \frac{P(X = k)}{M(X = k)} \\ &+ \textstyle\sum_{k\in\mathcal{X}} P(X = k) \cdot H_{P||M}(O|X = k) \\ &= H_{P||M}(X) + H_{P||M}(O|X) \end{aligned} \qquad (7)$$

where $H_{P||M}(O|X)$ and $H_{P||M}(X)$ are the conditional relative entropy and marginal relative entropy, respectively. Using the symmetric relationship we have that $H_{P||M}(O, X) = H_{P||M}(O) + H_{P||M}(X|O)$ where given that the conditional relative entropy is greater than zero, we have that:

$$H_{P||M}(O) \leq H_{P||M}(O|X) + H_{P||M}(X) \qquad (8)$$

If we consider O as the observable phenomenon, by denoting $f_k^1(o) = P(O = o|X = k)$, $f_k^2(o) = M(O = o|X = k)$, $c_k^1 = P(X = k)$ and $c_k^2 = M(X = k)$, the marginal distribution $P(O = o)$ and $M(O = o)$, $\forall o \in \mathbb{R}^n$ are equivalent to the mixture models $f^1(o) = \sum_{k=1}^{K} c_k^1 \cdot f_k^1(o)$ and $f^2(o) = \sum_{k=1}^{K} c_k^2 \cdot f_k^2(o)$, respectively. Consequently, the inequality presented in Eqn.(8) is equivalent to the previously presented upper bound proposed in [5], [8], given the fact that $D\left(f^1||f^2\right) = H_{P||M}(O)$.

Using this representation in terms of observable and non-observable random variables, we can explain mixture models as the marginalization of a joint phenomena because of the inability in observing one of the information sources. In this case, the difference between the proposed upper bound and $D\left(f^1||f^2\right)$ is given by how much discrimination information P has with respect to M in the process of using both random variables (observable and non-observable) as evidence. This interpretation is valid for the particular alignment given by the probabilistic mapping between O and X in the process of calculating $H_{P||M}(O, X)$, Eqn.(7), and consequently implies to have an additional knowledge of the underlying structure of the models to be compared; more precisely, in the characterization of the joint distribution $P(O, X)$ and $M(O, X)$. Given that this information is not available for mixture models, any joint P and Q distributions whose marginals are given by $f^1(o)$ and $f^2(o)$, respectively, characterize an upper bound for the KLD by Eqns.(7) and (8). Consequently, this problem can be seen as finding the optimal underlying injective alignment $\beta(\cdot)$ between the mixture components that minimize Eqn. (5), or equivalently the tightest upper bound. The criterion proposed in Eqn.(6) is as computationally implemented as a greedy approach, which approximates the optimal alignment under the asymptotic assumption formulated in [4], where mixture components tend to have the mentioned perfect intra-discrimination condition and both mixture models tend to define the same partition in the observation space $\mathbb{R}^n$. Furthermore, in this case the upper bound approximates the actual KLD between the models. The proof can be derived from (*Theorem 4*, [4]). This motivates in part the alignment criterion presented in the previous section, Eqn.(6).

IV. Experiments

In this section we evaluated the UB-KLD for left-to-right HMMs presented in section II, Eqn.(4), using the proposed upper bound for GMMs presented in section III. The first part is devoted to analyzing the correlation between the UB-KLD and a numerical estimation of the divergence using Monte Carlo simulation techniques, and the second part shows how a global indicator based on the UB-KLD correlates with the phone recognition accuracy (PAC), i.e. success in discriminating the basic phonetic units of a given language (40 in American English).

A. Speech Corpus and Experimental Set-up

The corpus used for this analysis provides speech data from children and adults grouped by age. The speech data (16Khz, NIST format) were obtained from 436 American English speaking children (ages 5-18 years), with an age resolution of 1 year. The database has 231 male speakers and 252 female speakers. This database comprises continuous read spoken commands and telephone numbers and has an average of 2300 utterances in each category. For the second part, section IV-C, the idea is to evaluate UB-KLD under a variety of different acoustic discrimination conditions by segmenting the data in age-dependent data groups (5-18 years) [15]. For each age category, a training and test set were generated using approximately 90% and 10% of the entire database, respectively.

In general we used context-independent HMMs with standard left-to-right topology, 3 inner observable states and GMMs with eight mixtures per state to model each of the 40 phones (acoustic realizations of the phonemes) of English. 39 dimensional feature vectors (13MFCCs, 13 delta, 13 acceleration coefficients) were calculated based on a 25msec Hamming window every 10msec.

Training step	UB-KLD	UB-KLDR	Mahalanobis
5	0.902	0.858	0.816
13	0.878	0.798	0.728
15	0.872	0.799	0.734
18	0.866	0.793	0.732
23	0.860	0.779	0.720

TABLE I

CORRELATION EVALUATION WITH NUMERICAL ESTIMATION OF THE DIVERGENCE

B. Correlation with Numerical Estimation of the Divergence

We evaluate the approximation goodness of the UB-KLD by means of correlating it with an estimation of the divergence using Monte Carlo simulation. In this scenario, we compare the UB-KLD with respect to two alternative HMM distance measures: the upper bound Kullback-Leibler Distance rate (UB-KLDR) proposed in [16], equivalent to the one proposed in [8], and the Mahalanobis distance presented in [17].

For this part, 40 mono-phone HMMs associated with the basic phonetic units were trained with the entire database (ages 5-18 years) using the HTK 3.0 toolkit and the standard maximum likelihood (ML) training strategy.Based on this set of HMMs the UB-KLD and divergence matrices were generated, in which each entry (i, j) of those matrices represent the symmetric extension of the UB-KLD, i.e. ($UB - KLD(\lambda^1||\lambda^2) + UB - KLD(\lambda^2||\lambda^1)$), and the divergence between the models λ^i and λ^j, respectively. In order to have an accurate numerical estimation of the divergence, more than 800 independent realizations of each of the mono-phone HMMs were generated, in which good convergence behavior of those estimations was achieved. Based on that, we computed the correlation coefficient between the symmetric extension of the UB-KLD and the divergence for each model i with respect to all the others, or in other words the i-row correlation between the UB-KLD and the divergence matrices. Finally, the average correlation coefficient was generated as a function of the frequency of the phonetic units in the training set. The same methodology was used to generate the correlation between the numerically estimated divergence matrix and the alternative distance measures UB-KLDR and Mahalanobis distance. The evaluation scenario considers models at different steps of the re-estimation process, Table I, to explore generality of the method.

Table I presents those results from which it is clear that the proposed UB-KLD significantly outperforms the two alternative distance measures in approximating the discrimination tendency of the divergence, in most of the cases 9% higher than the UB-KLDR and 18% higher than for the Mahalanobis distance — results representative of the other re-estimation scenarios that were evaluated. Figure 1 shows the row zero-mean unit-variance UB-KLD and divergence matrices for the iteration number 5 of the re-estimation process, in which it is possible to graphically see at a higher level of detail, the discrimination behavior of the UB-KLD relative to the divergence. This is a representative scenario of all the other evaluated experimental conditions presented in this section.

C. Application to Estimate Complexity of the Recognition Task

In this section we evaluate the applicability of the UB-KLD in providing a global discrimination indicator of the complexity of a recognition task [2]. For defining the indicator, we calculated the average pair-wise symmetric UB-KLD between all the mono-phone HMMs considering the frequency of those models in the training set. For this scenario context-independent HMMs were trained independently for each age group (ages 5-18). Again a total of 40 mono-phone HMMs were created in each of those data conditions.

This global upper bound Kullback-Leibler divergence (G-UBKLD) was calculated for all the age groups in our speech corpus, and across different phases of the training process. Simultaneously, a phone-recognition system was generated under all these different acoustic conditions and the recognition performance of this system was calculated independently for the training and test set mentioned in section IV-A. Table II shows the correlation coefficient between the proposed indicator (G-UBKLD) and the phone accuracy (PAC) in the training and test set, for some of the age dependent data group — results are representative of all the other data groups. In order to calculate this correlation coefficient we used the first seven consecutive steps in the training process for each data condition (age group) to calculate the G-UBKLD and the performance of the system in each of these steps.

These results show that this global discrimination indicator is highly correlated with the performance of the system and consequently, it can be used as a precise indicator of the average complexity of the inference task based on the acoustic information. Moreover, the results are consistent across different acoustic discrimination conditions (age dependent data groups). Fig. 2 shows the evolution of the normalized zero-mean and unit-variance G-UBKLD and PAC across the re-estimation process (training and test sets) for one representative experimental scenario, 8 years-old group. This figure graphically presents the close dependency between the G-UBKLD and the performance indicators (PAC) of the system.

In the general process of considering arbitrary number of re-estimation steps, the system will start presenting over-fitting effects to the training condition. In this scenario the global discrimination indicator (G-UBKLD) tends to be less correlated with the test PAC, while perhaps remaining highly correlated with the PAC of the training set, as expected. These results were observed in our experimental analysis but we do not present them.

V. CONCLUSIONS

This work formalized a criterion for defining a closed-form upper bound for the KLD between GMMs. The criterion is based on a similarity alignment between mixture components of the models, motivated by the asymptotic scenario in which this upper bound turns out to be actual KLD between the models. Moreover, information theoretic interpretations of this upper bound and the equivalent optimization problem were presented. Finally, this indicator was used to characterize an UB-KLD between standard left-to-right HMMs used in

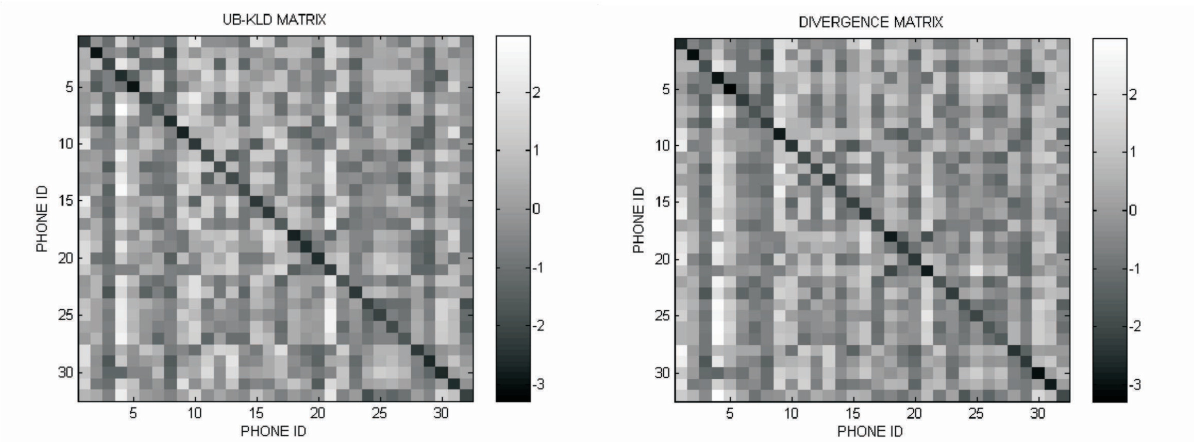

Fig. 1. Row zero-mean unit variance upper bound Kullback-Leibler divergence (UB-KLD) and Divergence matrices for acoustic-phonetic HMMs. Diagonal elements correspond to comparing same model (darker color implies smaller magnitude). Both graphics have the same magnitude scale, minimum equal to -3.2891 and maximum equal to 2.9792. Results obtained at iteration 5 of the Baum -Welch re-estimation algorithm.

Age Group	*Train. Set*	*Test Set*
5	0.9853	0.9202
6	0.9842	0.9860
8	0.9878	0.9735
9	0.9970	0.9839
10	0.9922	0.9863
11	0.9941	0.9924
12	0.9928	0.9845
13	0.9914	0.9577
16	0.9906	0.9642
17	0.9847	0.9755
18	0.9775	0.9681

TABLE II

CORRELATION BETWEEN THE GLOBAL UB-KLD AND PHONE ACCURACY

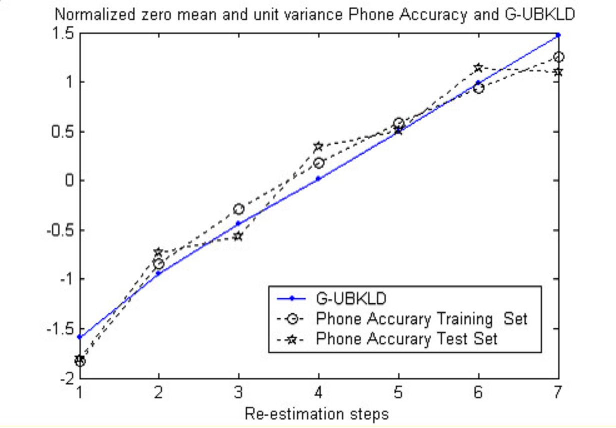

Fig. 2. Normalized zero-mean and unit-variance sequences associated with the Global UB-KLD, phone accuracy (PAC) in training and test set across 7 consecutive steps in the re-estimation process for the 8 year-old group.

speech recognition, where experimental evidence supports the goodness of this representation as a discrimination measure for HMMs.

ACKNOWLEDGMENT

This material is based upon work supported by awards from National Science Foundation, ONR-MURI,DARPA and the U.S. Army.

REFERENCES

[1] S. Kullback, *Information theory and statistics*, New York: Wiley, 1958.

[2] J. Silva and S. Narayanan, “Average divergence distance as a statistical discrimination measure for hidden Markov models,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 890–906, May 2006.

[3] M. N. Do and M. Vetterli, “Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden Markov model,” *IEEE Transaction on Multimedia*, vol. 4, no. 4, pp. 517–527, December 2002.

[4] N. Vasconcelos, “On the efficient evaluation of probabilistic similarity functions for image retrieval,” *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1482–1496, July 2004.

[5] Y. Singer and M.K. Warmuth, “Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy,” *in Advances in Neural Information Processing System*, vol. 11, pp. 578–584, 1998.

[6] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley Interscience, 1991.

[7] Y. Singer and M.K. Warmuth, “Training algorithm for hidden Markov models using entropy based distance functions,” *in Advances in Neural Information Processing System*, vol. 9, pp. 641–647, 1996.

[8] M. N. Do and M. Vetterli, “Fast approximation of Kullback - Leibler distance for dependence trees and hidden Markov models,” *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 115–118, April 2003.

[9] J. Silva and S. Narayanan, “An upper bound for the Kullback-Leibler divergence for left-to-right transient hidden Markov models,” *IEEE Transactions on Information Theory*, vol. submitted for review, June 2005.

[10] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[11] B. H. Juang and L. R. Rabiner, “A probabilistic distance measure for hidden Markov models,” *AT&T Technical Journal*, vol. 64, no. 2, pp. 391–408, 1985.

[12] Y. Ephraim and N. M. Merhav, “Hidden Markov processes,” *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1518–1569, June 2002.

[13] J.R. Norris, *Markov Chains*, Cambridge series in Statistical and Probabilistic Mathematics, 1999.

[14] F. Jelinek, *Statistical methods for speech recognition*, MIT Press, 1997.

[15] S. Yildirim and S. Narayanan, “An Information-Theoretic analysis of developmental changes in speech,” in *Proc. ICASSP*, April 2003.

[16] M. Vihola, M. Harju, P. Salmela, J. Suontausta, and J. Savela, “Two dissimilarity measures for hmms and their application in phoneme model clustering,” in *in Proc. ICASSP 2002*, 2002, pp. 933–936.

[17] M.-Y. Tsai and L.-S. Lee, “Pronunciation variations based on acoustic phonemic distance measures with applications examples of mandarin chinese,” in *in ASRU*, December 2003.