A Statistical Discrimination Measure for Hidden Markov Models based on Divergence

Jorge Silva, Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory, http://sail.usc.edu
Department of Electrical Engineering and Integrated Media Systems Center
University of Southern California Viterbi School of Engineering, **Los Angeles, CA 90089, USA**
jorgesil@usc.edu, shri@sipi.usc.edu

Abstract

This paper proposes and evaluates a new statistical discrimination measure for hidden Markov models (HMMs) extending the notion of divergence [1], a measure of average discrimination information originally defined for two probability density functions. The Average Divergence Distance (ADD) is proposed as a statistical discrimination measure between two HMMs, considering the transient behavior of these models. We show the analytical formulation of this discrimination measure, and demonstrate that this quantity is well defined for a left-to-right HMM topology with final non-emitting state, a standard model for basic acoustic units in Automatic Speech Recognition (ASR). Using experiments based on this discrimination measure, it is shown that ADD is a coherent way to evaluate the discrimination dissimilarity between acoustic models.

1. Introduction

Hidden Markov models (HMMs) have been highly successful in modeling complex time series phenomena such as speech. The need for comparing different HMMs, through appropriate distance measures, often arises in a variety of contexts. Consider for example automatic speech recognition. Some of the applications here include: evaluation of the re-estimation processes [2]; redefinition of acoustic units [3]; multilingual phoneme mapping [4]; vocabulary selection [5], and more recently in pronunciation variation analysis [6]. Furthermore, similar to the language-model perplexity measure [7], the HMM distance measure can be used to provide information about acoustic level complexity [8], in discriminating between acoustic models. The availability of quantitative indicators of complexity from models at different levels of linguistic abstraction in an ASR system can support more reliable information integration and help provide further insights about the role of each component in the overall decoding process.

The relative entropy or Kullback–Leibler distance (KLD), is the average discrimination information per observation between two hypotheses modeled as random variables [1]. In its original formulation these hypotheses are characterized by two probability density functions, *f1* and *f2*. The Kullback–Leibler distance of H1 respect to H2 is given by:

$$D_{KLD}(f_1\|f_2)=\int f_1(o)\log\left(\frac{f_1(o)}{f_2(o)}\right)\partial o \qquad (1)$$

Using this quantity it is possible to define the divergence $J(H_1,H_2)$ (Eqn. 2), a symmetric statistical measure of the discrimination information between two hypotheses [1]. Although it has been shown that $J(H_1,H_2)$ is not a distance in the space of probability distributions, this statistical measure is an accurate indicator of the complexity of discriminating between them, which explains its wide use in the context of discrimination and classification [2, 9, 10].

$$J(H_1,H_2)=D_{KLD}(f_1\|f_2)+D_{KLD}(f_2\|f_1) \qquad (2)$$

Despite its precise formulation in the context of random variables, there is not a closed form analytical expression for the divergence between hypotheses modeled as hidden Markov processes. Some previous efforts that have tried to extend the Kullback-Leibler distance concept to the case of HMMs focused just on the stationary behavior of the models [2, 9]. In general, these approaches are constrained to the case of ergodic Markov sources and they use their invariant distribution to asymptotically approximate discrimination indicators, the Kullback-Leibler distance rate (KLDR) in [2], and a deterministic upper bound of the KLDR in [9].

However, in speech recognition, the transient aspects of the models play a crucial role in the acoustic discrimination process. Hidden Markov models are used to incorporate the non-stationary behavior of basic phonetic or sub-word units in which the transient aspects of the model are reflected in the transient states, which introduce the statistical behavior of pseudo-stationary segments of the signal (20-30 ms) and the stochastic state-transition matrix, which represents the dynamic evolution of the process across its transient states. Moreover, if the invariant distribution for the classical left-to-right topology with a final non-emitting state is considered, this distribution puts all the probabilistic mass in the final state, disregarding all the relevant dynamic information of the process. Consequently, previous approaches to evaluating distance between two HMMs are not naturally extended to this context.

We address the problem of finding, and evaluating, a discrimination measure based on the divergence, considering the transient statistical behavior of HMMs, especially for ASR applications. We propose the Average Divergence Distance (ADD) as a statistical discrimination measure between HMMs. We show that the ADD is analytically well defined in our case of interest (left to right topology) and a coherent indicator of the discrimination information between acoustic units based on both theoretical analysis and experimental results. Also a dynamic programming technique is proposed to compute ADD recursively in the general case.

2. Average divergence distance (ADD)

Our method proposes to calculate an average discrimination indicator that considers the mapping between all the potential states of the HMMs being compared based on the well-defined divergence at the observation distribution level of the models.

First an elemental case is presented and then this result is extended to the general case, following the methodology proposed by Printz et. al in [8]. The elemental case considers that both models have only one potential hidden state alignment. Let us define models λ^1 and λ^2 both with a left-to-right topology, *N+1* states with final non-emitting state, and stochastic matrices with transition probability equal to 1. Under this restriction the KLD between the models is given by:

$$\begin{aligned} D_{KLD}\left(\lambda^1 \| \lambda^2\right) &= E_{\lambda^1}\left(\log \frac{P(o_1,...o_N/\lambda^1)}{P(o_1,...o_N/\lambda^2)}\right) \\ &= E_{\lambda^1}\left(\log \frac{P(o_1,...o_N/s_1,...s_N,\lambda^1)}{P(o_1,...o_N/s_1,..,s_N,\lambda^2)}\right) \\ &= E_{\lambda^1}\left(\sum_{i=1}^{N} \log \frac{P(o_i/s_i,\lambda^1)}{P(o_i/s_i,\lambda^2)}\right) = \sum_{i=1}^{N} D_{KLD}\left(P(o/s_i,\lambda^1)\|P(o/s_i,\lambda^2)\right) \end{aligned} \quad (3)$$

where $E(\cdot)$ refers to the expectation operator.

The first equality is because only a sequence of *N* observations can be generated, given the topology of λ^1 and the non-emitting final state. The second term is because only one state sequence is possible, given the assumption about stochastic matrices of λ^1 and λ^2. The third term is because of the Markov property, and the last expression is because of the KLD definition [1]. Using this result, the divergence $J(\lambda^1,\lambda^2)$ is the sum of the divergences between the observation distributions of the models under this particular state mapping, equation (4), below:

$$J\left(\lambda^1,\lambda^2\right) = \sum_{i=1}^{N} J\left(P(o/s_i,\lambda^1), P(o/s_i,\lambda^2)\right) \quad (4)$$

In this simple case, only one state mapping exists between the models, but in the general case multiple state associations are possible. As a consequence, the well-defined result obtained in Eqn.(4) can be naturally extended to define the Average Divergence Distance (ADD) as the expected value of the sum of the divergences $J\left(P(o/s_i,\lambda^1), P(o/s_j,\lambda^2)\right)$, given the underling state mapping process, that we denote by *S*, in Eqn.(5):

$$D^{ADD}\left(\lambda^1,\lambda^2\right) = E_{S}\left(J\left(\lambda^1,\lambda^2/S\right)\right) \quad (5)$$

It is important to note that $J\left(\lambda^1,\lambda^2/S=s\right)$ has the same structure presented in Eqn.(4), because it specifies the divergence that is constrained to only one particular state mapping. Therefore, $J\left(\lambda^1,\lambda^2/S=s\right)$ is a random variable of the underling state mapping process S, and consequently the ADD is the expected value of this random variable, per Eqn.(5). The process to derive this measure using dynamic programming techniques is presented in the next section by means of characterizing the state mapping process *S*.

2.1. State mapping process

Let us define the product Markov Chain $\lambda^{1\times2}$ modeling the statistical behavior of the state mapping process S. It is natural to assume that $\lambda^{1\times2}$ is the coupling of two independent Markov chains associated with the original HMMs that we want to compare. More precisely, consider the hidden Markov model $\lambda^1 = \left(\pi^1, A^1, B^1\right)$ defined in the state space S^1 and the hidden Markov model $\lambda^2 = \left(\pi^2, A^2, B^2\right)$ in the state space S^2, then the product Markov chain $\lambda^{1\times2} = \left(\pi^{1\times2}, A^{1\times2}\right)$ in $S^{1\times2} = S^1 \times S^2$ is defined as:

$$\pi^{1\times2}\left(s_1,s_2\right) = \pi^1{}_{s_1} \cdot \pi^2{}_{s_2}, \forall (s_1,s_2) \in S^{1\times2}$$

$$A^{1\times2}{}_{((s_1,s_2),(s'_1,s'_2))} = A^1{}_{(s_1,s'_1)} \cdot A^2{}_{(s_2,s'_2)}, \forall s_1,s'_1 \in S^1, \forall s_2,s'_2 \in S^2 \quad (6)$$

Let us consider that each state (s_1,s_2) in $S^{1\times2}$ has only one observation, the divergence between the observation distribution associated to this particular state mapping, $d_J(s_1,s_2) = J\left(P(o/s_1,\lambda^1), P(o/s_2,\lambda^2)\right)$, then using Eqn.(5) the Average Divergence Distance (ADD) in this context is given by:

$$D^{ADD}\left(\lambda^1,\lambda^2\right) = E_{\lambda^{1\times2}}\left(D_J(S)\right) = \sum_{s\in\left(S^{1\times2}\right)^N} D_J(s)\cdot p\left(s/\lambda^{1\times2}\right) \quad (7)$$

where $s = (s_n)_{n>0} \in \left(S^{1\times2}\right)^N$ represents a right-side state mapping sequence, and $D_J(s) = \sum_{n>0} d_J(s_n)$ is the divergence cost associated with the state mapping sequence s, extending the relation observed in Eqn.(4).

2.2. Dynamic programming implementation

The expression in Eqn.(7) considers the expected value of the entire possible path in the product Markov chain $\lambda^{1\times2}$. The structure of this problem is equivalent to the structure used in calculating the probability of an observation given a HMM [11], but with the additional complication that there is not restriction in the length of state mapping paths. Therefore, a dynamic programming approach can be used to solve this problem. The idea is to define the expected value of all the potential paths of length *N* constrained that the final state is a particular state s in $S^{1\times2}$, by the forward coefficient α_s^N, per equation (8):

$$\begin{aligned} \alpha_s^N &= E_{\lambda^{1\times2}}\left(\sum_{n=1}^{N} d_J(s_n), s_N = s\right), \forall s \in S^{1\times2}, \forall N > 0 \\ &= \sum_{s_1,s_2,...,s_{N-1}\in S^{1\times2}} \left(\left(\sum_{n=1}^{N-1} d_J(s_n) + d_J(s)\right)\cdot p\left(s_1,s_2,...,s_N = s/\lambda^{1\times2}\right)\right) \end{aligned} \quad (8)$$

The relationship between α_s^N and the original problem $D^{ADD}\left(\lambda^1,\lambda^2\right)$ is derived below:

$$D^{ADD}\left(\lambda^1,\lambda^2\right) = E_{\lambda^{1\times2}}\left(\sum_{n=1}^{\infty} d_J(s_n)\right)$$

$$D^{ADD}\left(\lambda^1,\lambda^2\right) = \lim_{N\to\infty} E_{\lambda^{1\times2}}\left(\sum_{n=1}^{N} d_J(s_n)\right)$$

$$D^{ADD}\left(\lambda^1,\lambda^2\right) = \lim_{N\to\infty} \sum_{s\in S^{1\times2}} \left(E_{\lambda^{1\times2}}\left(\sum_{n=1}^{N} d_J(s_n), s_N = s\right)\right)$$

And, finally from the definition of α_s^N

$$D^{ADD}\left(\lambda^1,\lambda^2\right) = \lim_{N\to\infty} \sum_{s\in S^{1\times2}} \alpha_s^N \quad (9)$$

Using the Markov property of the process it is possible to derive a dynamic programming recursion, forward equation, which allows solving the problem of length *N* as a function of the same problem of length *N-1*. The recursion is given below.

$$\alpha_s^N = \sum_{p \in S^{1\times 2}} \alpha_p^{N-1} \cdot A_{p,s}^{1\times 2} + d_J(s) \cdot p(s_N = s) \quad (10)$$

This dynamic recursion allows finding α^N for all N and consequently asymptotically approximates the ADD between the models (Eqn. 9). But it is necessary to analyze under what conditions α^N converges and, as a consequence, the ADD is well defined. The next section presents the results of this problem for the special case of HMMs with left–to-right topology with and a final non-emitting state.

2.3. ADD for left-to-right HMMs with final non-emitting state

Let us restrict the previous analysis to the case of HMMs with left-to-right topology and a final non-emitting state s_F. Also let us assume that each Markov model λ^1 and λ^2 starts in the initial state with probability one. Given that the number of observations between these two models should be the same, any transition to the set of state $S^F = \{(s_F^1, s), \forall s \in S^2\} \cup \{(s, s_F^2), \forall s \in S^1\}$ in $S^{1\times 2}$ represents the transition to the end of one of the models, where s_F^1 and s_F^2 are the final non-emitting state of λ^1 and λ^2 respectively. Because it is not possible to obtain the divergence if one of the states is non-emitting, it is natural to assume that:

$$d_J(s) = 0, \forall s \in S^F \subset S^{1\times 2} \quad (11)$$

Under this assumption, the divergence function $d_J(\cdot)$ is null for any state in S^F. Then S^F can be reduced to only one state representation s_F in the product state space $S^{1\times 2}$, the final absorbing state of null divergence function.

We can formally prove that in this case the ADD is well defined and it is given by the following expression.

$$D^{ADD}(\lambda^1, \lambda^2) = \lim_{N\to\infty} \sum_{n=1}^{N} \left(\sum_{j \in S_T} A_{j,s_F}^{1\times 2} \cdot \alpha_j^n \right) < \infty \quad (12)$$

where $S_T = S^{1\times 2} \setminus \{s_F\}$ is the collection of transient states, [12] (p. 24), in the product Markov chain $\lambda^{1\times 2}$. (The actual proof is outside the scope of this paper). This result shows that the ADD is a series in which its term, as a function of n, is a linear combination of forward coefficients of transient state at time n. Actually, we can show that the forward coefficients of transient states, α_s^n, are bounded by a geometric decay multiplied by a polynomial of finite order in n. This result is enough to prove that the series in Eqn. (12) converges and consequently, $D^{ADD}(\lambda^1, \lambda^2)$ is well defined.

In this expression, it is explicit that the ADD accumulates the discrimination information of the transient states of the models, i.e., it actually includes the set of states that present observations, and as a consequence are directly relevant in this discrimination analysis. Therefore, in its analytic formulation the transient behavior of the models is an explicit part of this discrimination measure. On the other hand, the geometric decay of the forward coefficient α_s^n in Eqn.(12) allows to efficiently approximate this infinite series with only a finite number of terms.

3. Experiments and discussion

For our experiments, we have chosen to use a corpus that provides speech data from both children and adults, separately grouped by age, to evaluate the proposed discrimination distance under different acoustic discrimination conditions. The speech data (16 kHz, NIST) were obtained from 436 American English speaking children (ages 5-18 years), with a age resolution of 1 year, and from 60 adult speakers (ages 25-50 years). The database has 261 male speakers and 282 female speakers. This database comprises read spoken commands and telephone numbers and has an average of 2300 utterances in each category. Different monophone HMM based recognition systems were created for each age dependent group (i.e., ages 5-18 and adults) using the HTK 3.0 toolkit and the training strategy proposed in [13]. Each HMM acoustic unit has a left-to-right topology with 3 inner states and 8 Gaussian mixtures per state. 39 dimensional features vectors (13MFCCs, 13 delta, 13 acceleration coefficients) were calculated based on a 25msec Hamming window every 10msec.

In this experiment, two results are invoked to calculate an upper-bound analytical expression for the KLD in the case of comparing two Gaussian mixture density functions, the upper bound approximation of the KLD between two mixture density proposed by Do et. al [9-10], and the well defined divergence measure for standard Gaussian distributions [1].

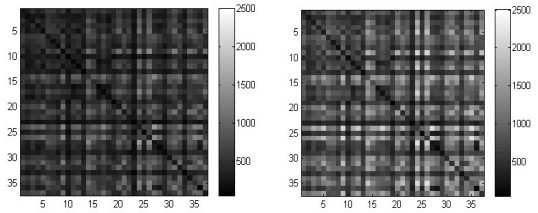

Figure 1: Graphical representation of ADD across monophone HMMs during training iteration steps 2 and 6. Diagonal elements correspond to comparing same models (darker color implies smaller divergence between the HMMs). Data from 5 years children were used for model training (re-estimation) shown in this example.

3.1. Acoustic discrimination evaluation

The ADD was calculated for each phone model with respect to all the others in the set. A graphical representation of this comparison matrix is shown in Figure 1 for two different stages in the training process. The data from five-year-old group is considered in this figure, but the results were found to be consistent across all the age groups in the database.

Figure 1 shows that the ADD has its minimum magnitude when the same HMMs are compared, as indicated by the diagonal component in the graphical representation, which is a desirable self-similarity property that any discrimination measure needs to posses. The analytical formulation of the ADD does not guaranty that this measure achieves its minimum when comparing a model with itself. Actually, it is possible that two different models with similar observation distribution and stochastic transition matrix generate this non-desirable condition. Nevertheless, in our experimental observations, such cases did not appear to be statistical significant, and even under the aforementioned condition, the self-ADD was the one that yielded the smaller magnitude.

The dynamic evolution of the ADD across the HMM training process is also presented in Fig. 1. The training process used was the standard maximum-likelihood technique based on the Expectation Maximization algorithm [14]. Clearly, this technique is not based on any discrimination principle, but it is expected to increase the discrimination capability of the models, if there is a reasonable separation among the hypotheses in the observation space. The evolution of the ADD for different stages of the training process, represented in two phases in Fig. 1, shows that the discrimination of the models based on the ADD, increases systemically in the re-estimation process, which is consistent with our previous observations. Based on these results it is reasonable to assume that ADD could give valuable information about the overall discrimination quality of the acoustic models. The next section emphasizes this point by means of generating a global indicator of the acoustic discrimination quality based on ADD and correlates this quantity with the classification-performance (Phone Recognition in this case).

3.2. Correlation of ADD with phone recognition performance

Let us define the relative ADD (RADD) of λ^1 respect to λ^2 as:

$$R^{ADD}(\lambda^1,\lambda^2) = D^{ADD}(\lambda^1,\lambda^2) - D^{ADD}(\lambda^1,\lambda^1) \qquad (13)$$

Then, if we take the mean of this quantity with respect to all of the models λ^2 , $E_{\lambda^2}\left(R^{ADD}(\lambda^1,\lambda^2)\right)$, we have an average normalized indicator of the ADD (RADD) for the model λ^1. This quantity is expected to be proportional to how easy it is to discriminate λ^1 with respect to all the other acoustic models. Then, taking the mean of $E_{\lambda^2}\left(R^{ADD}(\lambda^1,\lambda^2)\right)$ for all the models λ^1 , we get a global indicator of the complexity in discriminating basic acoustic units based on the ADD. We define this indicator as the Global Average Divergence Distance (GADD), Eqn (14).

$$G^{ADD} = E_{\lambda^1}\left(E_{\lambda^2}\left(R^{ADD}(\lambda^1,\lambda^2)\right)\right) \qquad (14)$$

We calculated GADD for different 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 calculated. Correlation coefficients between the proposed indicator (GADD) and the performance of the phone recognition, phone accuracy (PAC), were calculated for all the fifteen age groups. In order to calculate this correlation coefficient we used nine consecutive steps in the training process for each age group and the performance of the system in each of these steps was evaluated with the trained model set.

The correlation coefficient is over 97% in all the various data conditions. As a result, the GADD shows to be highly correlated with the phone-recognition performance under different acoustic conditions and across various stages of the training process. These results indicate that GADD can give a precise estimation of the acoustic discrimination complexity supporting the fact that the ADD is a coherent discrimination measure of the acoustic models.

4. Conclusions

We proposed the Average Divergence Distance, ADD, as a statistical discrimination measure for HMMs taking into consideration its transient behavior. This ADD measure considers the sub-stationary transient segment of the models, calculating the divergence in the entire potential mapping between the observation distribution of the models and the dynamic evolution, by means of weighting the cost of each alignment by its probability, using information obtained from the stochastic matrices of the models. The experimental evaluations presented in this paper show that the ADD is an accurate indicator of the discrimination behavior of basic acoustic models and can provide accurate information of the overall acoustic discrimination complexity in an ASR system.

5. Acknowledgements

The work was supported in part by NSF and by DARPA.

6. References

[1] S. Kullback, “Information Theory and Statistics”, New York: Wiley, 1958.

[2] 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.

[3] R. Singh, B. Raj., R. Stern, “Structured redefinition of sound units by merging and splitting for improved speech recognition”, in ICSLP, 2000.

[4] Kohler J., “Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds”, in ICSLP, 1996.

[5] P. Geutner, M. Finke, and A. Waibel, “Selection Criteria for hypothesis driven lexical adaptation”, in ICASSP, 1999.

[6] Ming–Yi Tsai and Lin – Shan Lee, “Pronunciation Variations Based on Acoustic Phonemic Distance Measures with Applications examples of Mandarin Chinese”, in ASRU December 2003.

[7] F. Jelinek, “Statistical Methods for Speech Recognition”, MIT Press, 1997.

[8] H. Printz and P. Olsen, “Theory and Practice of Acoustic Confusability”, in ISCA ITRW ASR2000, pp. 77—84, 2000.

[9] M. N. Do, “Fast approximation of Kullback – Leibler distance for dependence trees and hidden Markov models”, IEEE Signal Processing Lett. , vol. 10, No. 4, pp. 115-118, Apr. 2003.

[10] 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, Dec. 2002.

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

[12] J.R. Norris, “Markov Chains”, Cambridge series in Statistical and Probabilistic Mathematics, 1999.

[13] S. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland, “HTK book”, Cambridge Research Laboratory, 1997.

[14] A.P. Dempster, N. M. Laird, D.B. Rubin, “Maximum Likelihood Incomplete Data via EM Algorithm”, Journal of the Royal Statistical Society, Series B, Vol. 39, pp. 1-38, 1977.