TWO DISSIMILARITY MEASURES FOR HMMS AND THEIR APPLICATION IN PHONEME MODEL CLUSTERING

Matti Vihola[1]*, Mikko Harju*[2]*, Petri Salmela*[2]*, Janne Suontausta*[3] *and Janne Savela*[4]

Tampere University of Technology
[1]Institute of Signal Processing
[2]Institute of Digital and Computer Systems
P. O. B. 553, FIN-33100 Tampere, Finland
Email: matti.vihola@tut.fi

[3]Nokia Research Center
Speech and Audio Systems Laboratory
P. O. B. 100, FIN-33721 Tampere, Finland
[4]Department of Phonetics, University of Turku
FIN-20014 Turku, Finland

ABSTRACT

This paper introduces two approximations of the Kullback-Leibler divergence for hidden Markov models (HMMs). The first one is a generalization of an approximation originally presented for HMMs with discrete observation densities. In that case, the HMMs are assumed to be ergodic and the topologies similar. The second one is a modification of the first one. The topologies of HMMs are assumed to be left-to-right with no skips but the models can have different number of states unlike in the first approximation. Both measures can be presented in a closed form in the case of HMMs with Gaussian (single-mixture) observation densities. The proposed dissimilarity measures were experimented in clustering of acoustic phoneme models for the purposes of multilingual speech recognition. The obtained recognizers were compared to both recognition system based on previously presented dissimilarity measure and one based on phonetic knowledge. The performance of the multilingual recognizers was evaluated in the task of speaker independent isolated word recognition. Small differences were observed in the recognition accuracy of the multilingual recognizers. However, the computational cost of the proposed methods are significantly lower.

1. INTRODUCTION

The Kullback-Leibler (KL) divergence measure between HMMs cannot usually be presented in a closed form, so various approximations have been introduced [1, 2]. These approximations are based on Monte Carlo (MC) methods or simplifying approximations that lead to a closed form solution. In practice, the drawbacks of MC techniques are the extensive computational cost and slow convergence properties. On the other hand, the closed form approximations presented are limited to very specific class of HMMs, e.g. HMMs with discrete observation densities [2]. This paper introduces two approximations that can be represented in a closed form for HMMs with Gaussian observation densities.

The dissimilarity measures between HMMs can be utilized e.g. in model selection and clustering [1]. Parameter tying techniques have been used in order to reduce the total number of parameters in a speech recognition system. The tying techniques take place also when training data is insufficient or unavailable. This concerns especially embedded systems where the resources, e.g. memory, are very limited. Model-level tying of parameters has been applied in training of multilingual phone models [3].

The organization of this paper is as follows. In Section 2, the Kullback-Leibler divergence is defined and in addition, a brief overview is given of the previously presented approximations of the divergence in the case of HMMs. The two proposed approximations are introduced in the end of Section 2. In Section 3, the experimental setup including speech recognition systems, speech corpuses, unseen languages and the model clustering framework, are explained. Section 4 gives an overview of the obtained results.

2. DISSIMILARITY MEASURES

A well-known dissimilarity measure between two probability distributions is the Kullback-Leibler divergence. It characterizes the discriminating properties of two probabilistic models λ and ξ, and is defined as [4]

$$J(\lambda, \xi) = I(\lambda : \xi) + I(\xi : \lambda) \tag{1}$$

where

$$I(\lambda : \xi) = E_\lambda \left\{ \log \frac{p_\lambda}{p_\xi} \right\} \tag{2}$$

is the directed divergence from λ to ξ. In Equation 2, the expectation E_λ is taken with respect to the distribution of the model λ, and the probability density functions of the corresponding models are denoted by p_λ and p_ξ. All the dissimilarity measures presented below are directed divergences. The corresponding symmetric dissimilarity measures are obtained using Equation 1.

2.1. Previous approximations of the KL-divergence measure

Juang et al. studied the KL-divergence between HMMs using MC simulations. The models were assumed ergodic, and the measure was defined as the mean divergence per observation sample [1]. The measure was given as

$$\widehat{I}_{MC}(\lambda : \xi) = \frac{1}{T} \log \frac{p_\lambda(O^\lambda)}{p_\xi(O^\lambda)} \tag{3}$$

where O^λ is an observation sequence generated by model λ. The length of the sequence is T, and the likelihoods given the models λ and ξ are $p_\lambda(O^\lambda)$ and $p_\xi(O^\lambda)$, respectively. The method is valid for HMMs with arbitrary observation probability distributions [1].

Köhler presented a variant of the measure given in Equation 3. Tokens extracted from speech corpus were used in the evaluation

Submitted to *ICASSP2002, scheduled for May 13-17, 2002 in Orlando, Florida, USA.*

of the measure instead of a generated random sequence. Furthermore, the models were not ergodic as in Equation 3, but left-to-right phone models. Estimate of the divergence was defined as a sample mean of the log-likelihood differences over the tokens

$$\widehat{I}_{SP}(\lambda:\xi) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\log\frac{p_\lambda(O_i^\lambda)}{p_\xi(O_i^\lambda)} \tag{4}$$

where O_i^λ is the i:th token of length T_i corresponding the model λ.

Some approximations of the measure in Equation 3 were proposed and compared by Falkhausen et al. [2]. The assumed ergodicity property of the originally left-to-right models was achieved by substituting the transitions to non-emitting state with transitions to the first emitting state. The first approximation, denoted by I^*_{MC}, used the likelihood of the most likely state sequence. This means that the likelihoods $p(O)$ in Equation 3 were substituted with likelihoods of the most likely state sequences $p^*(O) = \max_Q p(O,Q)$. Only minor differences were observed when comparing the behavior of I_{MC} and I^*_{MC} [2]. The second approximation proposed assumed discrete observation density HMMs with similar topologies [2]. Moreover, the most likely state sequences of both models were assumed to be equal with the state sequence that generated the token O^λ, i.e. $Q = Q^*_\lambda = Q^*_\xi$. It is straightforward to evaluate the obtained measure, as no Monte Carlo simulations are needed anymore. The resulting approximation can be written in closed form as [2]

$$\begin{aligned}\widehat{I}(\lambda:\xi) =& \sum_i r_i \sum_j a^\lambda_{ij}\log\left(a^\lambda_{ij}/a^\xi_{ij}\right)\\ &+\sum_i r_i\sum_k b^\lambda_{ik}\log\left(b^\lambda_{ik}/b^\xi_{ik}\right)\end{aligned} \tag{5}$$

where $A = [a_{ij}]$ is the $N\times N$ transition matrix and $B = [b_{ik}]$ the $N\times M$ observation probability matrix. The steady-state probabilities r_i of the ergodic Markov chain of model λ are solved from $\boldsymbol{r}^T = \boldsymbol{r}^T A^\lambda$ with constraint $\sum_i r_i = 1$.

2.2. Proposed approximations

The approximation in Equation 5 assumed discrete observation densities. However, the last sum term in Equation 5 is the directed divergence between the observation distributions of state i of the models λ and ξ. Thus we can write the Equation 5 in the form

$$\begin{aligned}\widehat{I}_1(\lambda:\xi) =& \sum_i r_i \sum_j a^\lambda_{ij}\log\left(a^\lambda_{ij}/a^\xi_{ij}\right)\\ &+\sum_i r_i I(b^\lambda_i : b^\xi_i)\end{aligned} \tag{6}$$

where b^λ_i and b^ξ_i are the observation probability distributions of the models λ and ξ in state i, respectively. Now the only terms that are dependent of the observation probability distributions are the directed divergences between the corresponding distributions of the states, $I(b^\lambda_i : b^\xi_i)$. The Equation 6 generalizes the Equation 5 for arbitrary observation densities. The other simplifying assumptions made in deriving the approximation in Equation 5 remain. In the case of Gaussian observation distributions, the cross-state directed divergences can be expressed in a closed form as [4]

$$\begin{aligned}I(b_1:b_2) = \frac{1}{2}\Big[&\log\frac{|\Sigma_2|}{|\Sigma_1|} + \operatorname{tr}\left(\Sigma_1(\Sigma_2^{-1}-\Sigma_1^{-1})\right)\\ &+\operatorname{tr}\left(\Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T\right)\Big]\end{aligned} \tag{7}$$

where Σ and μ denote the covariance matrices and the mean vectors of the distributions, respectively.

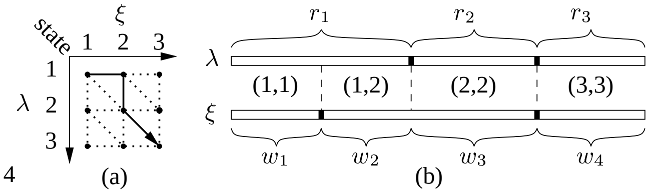

Fig. 1. The possible state alignments of the models λ and ξ are shown as dotted lines in (a). The solid line corresponds the alignment of the models shown in (b).

The assumption made in deriving the approximation in Equation 6 is that the most likely state sequences of both of the models are equal to the state sequence of the generating model. The assumption can be considered realistic with the model λ. However, there is no reason to assume that the model ξ would follow the same sequence. Let us draw one sample from the observation distribution of i:th state of the model λ. It is easy to agree that in most of the cases the likelihood of the generating distribution is greater than likelihoods given by other distributions. This justifies why the most likely state sequence of model λ should correspond the generating sequence in average. In the case of the other model ξ, the best likelihood is most likely given by the state-dependent density closest to the corresponding density of λ. This gives us a reason to find such a state alignment between the models λ and ξ that the states with best matching distributions coincide.

Assume now that the topology of both HMMs is left-to-right. The transitions are either self-transitions or transitions to the successor state. We introduce now a set of possible state alignments $\mathcal{Q}$ between the models λ and ξ. The alignments are illustrated in Figure 1a in the case of HMMs with three emitting states. The alignments are restricted such that the start and end points of the models are fixed, and occur at the same time. The transitions of model ξ can drift under the restriction that the segmentation of one state of λ can be divided into segments of equal duration. A new measure can be obtained for this kind of left-to-right proceeding models in the following way. The divergence measure is evaluated for each possible alignment in a similar way to Equation 6, and the alignment is picked that produces the minimum value. The reasoning above justifies this, as the divergence measures between the states are minimized when the corresponding observation distributions are most similar. The measure can be expressed as

$$\begin{aligned}\widehat{I}_2(\lambda:\xi) = \min_{(\boldsymbol{q},\boldsymbol{s})\in\mathcal{Q}}\Bigg\{&\sum_{i=1}^{L-1} w_{q_i}\Bigg[a^\lambda_{q_iq_i}\log\frac{a^\lambda_{q_iq_i}}{a^\xi_{s_is_i}}\\ &+a^\lambda_{q_iq_{i+1}}\log\frac{a^\lambda_{q_iq_{i+1}}}{a^\xi_{s_is_{i+1}}} + I(b^\lambda_{q_i}:b^\xi_{s_i})\Bigg]\Bigg\}\end{aligned} \tag{8}$$

where $\mathcal{Q}$ is the set of all possible alignments $(\boldsymbol{q},\boldsymbol{s})$. The models λ and ξ are at states q_i and s_i, respectively, at step i. The last, non-emitting states of the models are appended to the state vectors $\boldsymbol{q}$ and $\boldsymbol{s}$. The length of the alignments is $L = \dim \boldsymbol{q} = \dim \boldsymbol{s}$. The weight vector $\boldsymbol{w}$ is defined as $w_i = r_i/c_i$ where c_i denotes the count of state i in state vector $\boldsymbol{q}$. For example, the values corresponding the case in Figure 1b are $\boldsymbol{q} = (1, 1, 2, 3, 4)$, $\boldsymbol{s} = (1, 2, 2, 3, 4)$ and $\boldsymbol{w} = (r_1/2, r_1/2, r_2, r_3)$.

The approximation in Equation 8 can be evaluated for HMMs with different number of states unlike the approximation in Equation 6. In both Equation 6 and 8 the directed divergence measure between state-dependent observation densities can be expressed as in Equation 7 if the observation densities are Gaussian.

3. EXPERIMENTAL SETUP

The dissimilarity measures were experimented in the task of acoustic phoneme model clustering in the training of multilingual speech recognition system. The performance of the resulting speech recognition systems was evaluated in speaker independent isolated word recognition task. SpeechDat(II) databases of the corresponding languags were used in both training and testing of the recognition systems [5]. The test set consisted of 4000 isolated words spoken by total of 1000 speakers. The test vocabulary for each language had approximately 200 words including application words, isolated digits and forename surname combinations. During recognition, only the vocabulary of the target language was set active. The training set consisted of 4000 phonetically rich sentences for each language. The test and training sets shared no common speakers. The train and test sets have been described in more detail in [6].

The language dependent (LD) baseline recognition systems were trained for the seven languages shown in Table 3. These recognition systems consisted of 31 to 51 monophone HMMs. Two extra models were used to model silence and short pause. The short pause model was a single-state model tied to center state of silence model. All the other models had three emitting states. The emission probability density functions were mixtures of eight Gaussian densities with diagonal covariance matrices. The models were trained using the embedded Baum-Welch re-estimation from flat start [7]. The multilingual (ML) recognition systems were trained using the training data sets for the five source languages: German, English, Spanish, Finnish and Italian. The training procedure was identical to the one of the LD recognizers. The label files and the dictionary were changed to correspond the phone clusters of each system.

The speech material was parametrized using a mel frequency cepstral coefficient (MFCC) front-end. Feature vector consisted of 13 cepstral coefficients of which the zeroth, C_0, was replaced with frame energy. The first and second time derivatives of the elements were appended to the feature vector. Mean normalization was applied to the elements of the feature vector [8]. In addition, the variance of the energy coefficient and its derivatives were normalized as described in [8].

3.1. Knowledge-based ML recognition systems

The multilingual recognition system SAMPR was based on the method IPA-MAP described in [3]. The phonemes of different languages were clustered according to their phonetic symbol. The derived SAMPR system had a total of 105 multilingual phone models corresponding all the unique SAMPA symbols present within the databases of the five source languages.

Table 1. Mapping of the new phonemes of the two unseen languages. In Swedish the new retroflex consonant phonemes were split into separate phones, e.g. /rn/ → /r/ /n/.

French			
/e˜/	→	German	/E/ /N/
/9˜/	→	German	/9/ /N/
/R/	→	German	/r/
/H/	→	Finnish	/y/

Swedish			
/9:/	→	German	/9/
/{:/	→	Finnish	/{{/
/}:/	→	English	/u:/
/u0/	→	English	/U/

The knowledge-based recognition system, referred to as KB, was obtained with straightforward simplifications of the phoneme cluster definitions of SAMPR. The recognition system had no explicit models for long vowels and double consonants. Such phones were replaced with two single ones, e.g. /e:/ and /ee/ → /e/ /e/. In addition, the geminate affricates in Italian language were replaced with the preceding plosive and the following affricate, e.g. /ddz/ → /d/ /dz/. This way the total number of phone models was reduced to 64 from the starting point of 105.

3.2. ML systems based on dissimilarity measures

The clustering based on dissimilarity measures was performed using agglomerative clustering algorithm with complete linkage criterion [9]. The symmetric dissimilarity measures $\widehat{J}$ were obtained according to Equation 1 from the corresponding directed measures $\widehat{I}$. The measures $\widehat{J}_1$ and $\widehat{J}_2$ were used. In addition, the measure $\widehat{J}_{SP}$ was used, but the full likelihoods were approximated with the likelihoods of the most likely state sequences. This measure is referred to as SP1. These measures were evaluated using phoneme models with Gaussian densities. These single-mixture models were derived from the training procedure of the LD recognizers, just before splitting the Gaussian observation densities into mixtures of Gaussians. The SP dissimilarity measure was obtained also with the final LD phone models with 8 mixtures. This measure is referred to as SP8.

The estimates of the SP measures were obtained using up to 1000 tokens per LD phoneme model. These tokens were extracted from training speech material according to phone level segmentation of the utterances. The segmentation was performed using LD recognition systems. The phoneme models with insufficient data for estimating the measures (under 50 tokens) were not included in the clustering framework. Instead, they were assigned manually to the best matching cluster. In addition, the clustering was constrained such that phoneme models of the same language were not allowed in a same cluster. This procedure was applied identically also with the dissimilarity measures $\widehat{J}_1$ and $\widehat{J}_2$.

3.3. The unseen languages

There were two unseen languages, French and Swedish, included in the tests. They are unseen, for no data from these languages was used in training of any of the multilingual recognizers. The tying of the phonemes of these languages was based on phonetic knowledge. Each phoneme was tied explicitly to one language dependent phoneme model of the five source languages, e.g. French /@/ → English /@/. Then, before performing the test, each phoneme was mapped to a multilingual phone model accordingly. This tying was the same within all the ML recognition systems evaluated. The ty-

Table 2. Normalized values of dissimilarity measures between English /i:/ and a set of models.

Dissimilarity	English				German				Spanish		
measure	/I/	/eI/	/e/	/m/	/i:/	/I/	/e:/	/m/	/i/	/e/	/m/
SP8	0.2165	0.2957	0.8355	0.9120	0.2319	0.2975	0.3717	0.9036	0.1887	0.4778	1.0398
SP1	0.2051	0.2787	0.8727	0.8973	0.2433	0.2295	0.3026	0.9836	0.1939	0.4406	1.1032
$\widehat{J}_1$	0.2335	0.3077	0.7803	0.7496	0.3190	0.1962	0.1929	0.8172	0.3258	0.3784	0.9694
$\widehat{J}_2$	0.2993	0.3771	0.8730	0.8934	0.4067	0.2381	0.2473	0.9579	0.3568	0.4777	1.0932

Table 3. Average word recognition rates.

Recognition	Number of	Languages in multilingual training set						Unseen languages		
system	phones	German	English	Spanish	Finnish	Italian	Avg.	French	Swedish	Avg.
LD	303	83.32	78.40	95.78	95.35	92.07	88.98	82.77	85.29	84.03
SAMPR	105	79.60	63.38	94.55	92.30	93.18	84.60	49.73	69.52	59.62
KB	64	79.07	59.67	93.22	91.07	92.68	83.14	50.03	66.27	58.15
SP8	64	76.85	64.53	93.30	90.55	92.10	83.47	54.80	65.54	60.17
SP1	64	80.12	63.00	93.00	90.97	92.00	83.82	56.50	68.82	62.66
$\widehat{J}_1$	64	78.10	65.75	91.28	89.40	89.10	82.73	46.72	66.54	56.63
$\widehat{J}_2$	64	78.60	64.43	91.45	90.05	91.75	83.26	54.77	66.68	60.73

ing of the SAMPA symbols not present in the multilingual system is shown in Table 1.

4. RESULTS

In Table 2, the normalized values of the dissimilarity measures are shown between English /i:/ and a set of phones from three languages. The normalization applied is such that the mean dissimilarity between two models is one. There was no notable difference between the normalized dissimilarity values obtained using Gaussian (SP1) or eight-mixture Gaussian (SP8) observation distributions. The minimum dissimilarity value was achieved between English /i:/ and Spanish /i/ when measures based on speech data (SP) were used. German /e:/ and /I/ were the best matching phones when the approximations $\widehat{J}_1$ and $\widehat{J}_2$ were used, respectively. The most dissimilar model from the set shown was Spanish /m/, according to all the measures evaluated.

The recognition performances of the LD and ML speech recognition systems are shown in Table 3. The degradation from language dependent recognition systems into 105 model SAMPR system was moderate within the languages in the training set of the ML recognizers. The further reduction of the performance was small when the number of models was dropped into 64 in the KB system. The recognition systems based on computational model tying ($\widehat{J}_1$, $\widehat{J}_2$ and SP) can be considered comparable to the knowledge based (KB) system. The system based on approximation $\widehat{J}_2$ has comparable performance to the KB and SP systems. The recognition results of the unseen languages were significantly lower compared to the corresponding LD systems, especially for French.

5. CONCLUSIONS

We have presented two approximations of the Kullback-Leibler divergence measure for HMMs, and experimented them in the task of acoustic model tying. The approximations can be presented in closed form in the case of Gaussian observation densities. The proposed measures showed to be applicable in the task of clustering acoustic phoneme models, and only minor differences were observed in recognition performance compared to previously presented techniques. The computational costs of the proposed dissimilarity measures are significantly lower than of the measures estimated from speech tokens or generated random sequences.

6. REFERENCES

[1] 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.

[2] M. Falkhausen, H. Reininger, and D. Wolf, “Calculation of distance measures between hidden Markov models,” in *Proceedings of Eurospeech ‘95*, Madrid, 1995, pp. 1487–1490.

[3] J. Köhler, “Multilingual phone models for vocabulary-independent speech recognition tasks,” *Speech Communication*, vol. 35, no. 1-2, pp. 21–30, Aug. 2001.

[4] S. Kullback, *Information Theory and Statistics*, Dover Publications, New York, 1968.

[5] R. Winski, “SpeechDat: Definition of corpus, scripts and standards for fixed networks,” Tech. Rep. LE2-4001-SD1.1.1, http://www.speechdat.org/, Jan. 1997.

[6] M. Harju, P. Salmela, J. Leppänen, O. Viikki, and J. Saarinen, “Comparing parameter tying techniques for multilingual acoustic modelling,” in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, Sept. 2001, pp. 2729–2732.

[7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*, Cambridge University Engineering Department, July 2000.

[8] O. Viikki and K. Laurila, “Cepstral domain segmental feature normalization for noise robust speech recognition,” *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, Aug. 1998.

[9] S. Theodoridis and K. Koutroubas, *Pattern Recognition*, Academic Press, San Diego, 1999.