A Study on Separation between Acoustic Models and Its Applications

Yu Tsao, Jinyu Li and Chin-Hui Lee

School of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
{yutsao, jinyuli, chl}@ece.gatech.edu

Abstract

We study separation between models of speech attributes. A good measure of separation usually serves as a key indicator of the discrimination power of these speech models because it can often be used to indirectly determine the performance of speech recognition and verification systems. In this study, we use a probabilistic distance, called generalized log likelihood ratio (GLLR), to measure the separation between a model of a target speech attribute and models of its competing attributes. We illustrate five applications to compare separations among models obtained over multiple levels of discrimination capabilities, at various degrees of acoustic definitions and resolutions, under mismatched training and testing conditions, and with different training criteria and speech parameters. We demonstrate that the well-known GLLR distance and its corresponding histograms also provide a good utility to qualitatively and quantitatively characterize the properties of trained models without performing large scale speech recognition and verification experiments.

1. Introduction

In real-world pattern matching problems, such as automatic speech recognition (ASR) [1] and utterance verification (UV) [2], the true distributions of the patterns to be matched are often not precisely known. Thus the performance of such systems are usually determined by running experiments over a representative collection of evaluation samples intending to cover all possible variations of testing conditions using models created in a separate training phase. In many cases such an endeavor can be very challenging, if not impossible, in order to collect a large enough testing set that will produce statistically significant results. We are therefore interested in developing techniques that can be used to estimate the performance and behavior of real-world systems without conducting large scale experiments. Intuitively the separation between competing models in the same system serves as an important indicator to accomplish such purposes. For example model-based error estimation algorithms have been shown capable of predicting ASR performance [3].

Learning from minimum classification error (MCE) [4] and minimum verification error (MVE) [5] training formulations, the misclassification measure provides a quantitative indicator to represent a distance between a target model and its competing models. It can be used to measure the model separation as well. MCE and MVE can then be considered as a way to find model parameters that enhances the overall separation of the collection of models. A closer look at the misclassification measure reveals that it can also be considered as a probabilistic distance, called *generalized log likelihood ratio* (GLLR), commonly used in statistical hypothesis testing [6], if a log likelihood function is used to compute the class discriminant function [7]. GLLR also plays a key role in evaluating speech attribute detectors in a new speech research paradigm we are currently exploring under the ASAT (automatic speech attribute transcription) project [8]. In this study we illustrate a number of applications of the GLLR measure, and demonstrate that GLLR provides a good utility to characterize the discrimination capabilities of trained models without running large scale ASR and UV experiments.

2. Characterization of Model Separation

We now discuss issues related to computing GLLR measures and show that the corresponding histograms obtained from the sample GLLR values of target and non-target sets serve as useful tools to visually analyze model separation, and predict system performance for many ASR and UV tasks.

2.1. Defining target and competing sets

In pattern verification of a signal X, we first define a null hypothesis, H_0, and an alternative hypothesis, H_1, with H_0: {X is generated from S_0} versus H_1: {X is generated from any source but S_0}. A statistical test is then designed to divide the signal space S_X into two complimentary regions such that we reject hypothesis H_0, if $X \notin S_0$, and accept H_0, if $X \in S_0$. A tutorial can be found in [9].

In speech problems, H_1 is usually a composite hypothesis consisting of many signal classes. It has been shown that only the most competitive classes to H_0 need to be considered. This is usually accomplished by finding a speaker or phone "cohort" set [10, 2]. In this study, the cohort set is determined by selecting models that obtained the highest likelihood values when evaluating training data from the target class.

2.2. Computing target and competing scores

The LLR measure used in verification problems is defined as:

$$\mathrm{T}(X \mid \lambda_0, \lambda_1) = \log[\ell(X \mid \lambda_0)] - \log[\ell(X \mid \lambda_1)]. \quad (1)$$

where λ_0 and λ_1 are the parameters for the target model and non-target model, with $\log[\ell(X \mid \lambda_0)]$ and $\log[\ell(X \mid \lambda_1)]$ representing the target and competing scores, respectively.

When we use a cohort set for the target to calculate the non-target score generated by multiple competing models, the modified LLR score in Eq. (2) is called a generalized log likelihood ratio (GLLR) computed as follows:

$$\mathrm{T}(X \mid \lambda_q, \overline{\Lambda}_q) = \log[\ell(X \mid \lambda_q)] - \log[f(X \mid \overline{\Lambda}_q)] \quad , \quad (2)$$

where λ_q is a model for the target q, and $\overline{\Lambda}_q$ represents the set of competing models. The second term in the right hand side of Eq. (2) is an L_η norm of the scores in the cohort set C_q with size $|C_q|$ of the claimed target q. Eqs. (2) and (3) are commonly used in MCE algorithm [4]:

$$f(X \mid \bar{\Lambda}_q) = \{ \mid C_q \mid^{-1} \sum_r \exp[\eta \log \ell(X \mid \lambda_r)] \}^{1/\eta} \quad (3)$$

2.3. Preparing competing GLLR histograms

Based on the GLLR scores evaluated on samples of target and non-target segments in a set of speech utterances, a pair of GLLR histograms can be obtained with Eqs. (2) and (3). Figure 1 is an example of a typical GLLR plot with the right distribution (or histogram) curve representing the samples from the target source ($X \in S_0$), and the left curve depicting the sample distribution of the non-target source ($X \notin S_0$). The shaded region to the left of the vertical threshold line under the target curve gives the Type I error which is target samples missed. On the other hand, the shaded region to the right under the non-target curve represents the false alarms in detection. The smaller the regions the less the errors will be. Therefore, the performance of verification or recognition systems with the given models can be predicted (e.g. [3]). It is clear that this set of GLLR histograms can be generated for any verification problems we are interested in ASR and UV.

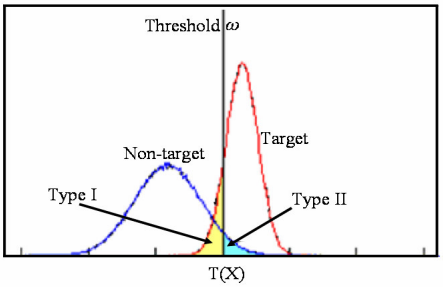

Figure 1: An illustration of GLLR plot for pattern verification

2.4. GLLR as a measure of separation

It is noted that GLLR is also a good measure for estimating the separation between a target and its competing cohort models. Therefore, it is easy to visually analyze the separation between two sets of models by examining the GLLR plots. New training and compensation algorithms can be developed to move the target and non-target curves. The effectiveness of different speech parameters, speech attributes, or model resolutions can be evaluated by comparing the overlap regions for each case. By moving the right curve to the right, or the left curve to the left or both, it is clear that it results in more separation between the two sets of competing models. It also indicates reduced Type I and Type II errors. Since minimizing errors and maximizing the model separation are closely related, it is clear to see why MCE and MVE algorithms have been shown very effective in many ASR and UV applications.

3. Applications of Model Separation Measures

In this following, we illustrate five applications of GLLR to compare separation among models obtained over multiple levels of discrimination capabilities, at various levels of acoustic definitions and resolutions, under mismatched training and testing conditions, and with different training criteria and speech parameters. We show that the GLLR separation measures and their corresponding histograms are good utilities to quantitatively and qualitatively study the properties of trained models without carrying out an extensive set of ASR and UV experiments.

In all the following experiments, both TIMIT and NTIMIT (Network TIMIT) databases [11] are used. Data in TIMIT were recorded with high-quality desktop microphones in a clean environment at a 16 KHz sampling rate. Excluding the speech materials reserved for speaker adaptation, there are 3696 and 1344 utterances in the standard training and testing sets, respectively. The NTIMIT data were obtained by passing the TIMIT version over dial-up lines, intending to simulate channel and noise distortion over the telephone network.

We used the entire training sets in the TIMIT database to train hidden Markov models (HMMs) [12] for phones and speech attributes. All HMMs were either related to a set of 45 English phones or another set of five manners of articulation, namely vowel, fricative, stop, nasal and approximant [13], plus silence. Almost all models have 3 states with each state characterized by a mixture Gaussian density with 8 mixture components. In most cases we used a feature vector of 39 elements, consisted of 13 MFCC parameters plus their first and second time derivatives, a commonly adopted feature vector used in most state-of-the-art ASR systems (e.g. [1]).

3.1. Model separation and acoustic discrimination

First we are interested in any correlation between model separation and acoustic discrimination capabilities. Two vowels, /ix/ (in tension) and /ay/ (in sunshine), were chosen for illustration. We used the five most competitive phones for /ay/, namely {/ah/, /aa/, /ae/, /eh/, /ao/} obtained from recognition results over the training set, to form its corresponding cohort set. Similarly, the five most competitive phones, {/ih/ (in shinbone), /ax/, /eh/, /uw/, /uh/}, to /ix/ were used to build the cohort set for /ix/. Based on some phonetic knowledge, the diphthong /ay/ is usually considered easier to recognize than /ix/, so the separation of /ay/ from other competing sounds is expected be larger than that of the phone /ix/ from its competing sounds. Figure 2 validates our assumption. It is seen that the overlap region in the top panel for /ix/ is clearly larger than that in the bottom panel for /ay/. This utility can be used to compare the degree of difficulty in recognizing and verifying different phones. We can also use the cohort set for each phone to evaluate the confusability of competing words in an ASR vocabulary, and try to avoid confusable pairs as much as possible in vocabulary design.

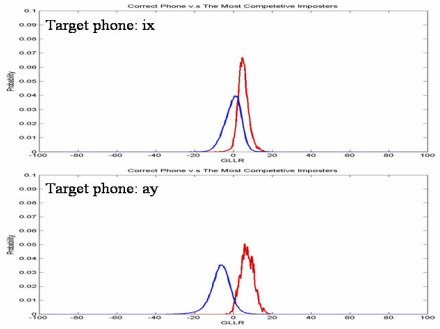

Figure 2: Model separation and acoustic discrimination

Another way to examine the properties of the model separation measure is to list recognition errors as shown in Table 1. Since Figure 2 indicates that there are much more

Type II errors for phone /ix/ when compared to phone /ay/, we predict that sound /ix/ is easier than sound /ay/ to be substituted by other competitive sounds. The results from Table 1 confirm the information displayed in Figure 2.

Table1: Errors for two phone models /ay/ and /ix/

Phone model	/ay/	/ix/
Correct	77.37%	40.11%
Substitution	18.64%	41.96%
Deletion	3.99%	17.93%
Insertion	6.07%	3.84 %

3.2. Model separation and acoustic mismatch

Next we are interested in comparing model separation in mismatched conditions. Phone models built from the TIMIT database were used, and both the testing sets from TIMIT and NTIMIT databases were collected to make GLLR plots for comparison. Since the spectral contents in the higher frequency bands have been removed in the telephone data, it is expected that the discrimination among fricative sounds is likely to be seriously degraded, more than the vowel sounds.

In Figure 3, we compare vowel /iy/ (in sheet) with fricative /sh/ (in sheet). The two plots in the top panels display results for matched testing conditions. They clearly show that the fricative /sh/ is easier to recognize than the vowel /iy/. When the testing data were from the mismatched NTIMIT database, it is noted that the overlap region to discriminate /sh/ is significantly increased in the bottom right panel, while the increase for /iy/ in the bottom left panel was not as serious. This validates our assumptions that for phone /sh/, the separation between the target and its competing models will be significantly reduced in a mismatched environment, and it is believed that the recognition performance will also be greatly degraded. On the other hand, the separation for the vowel phone /iy/ does not change as much in mismatched conditions.

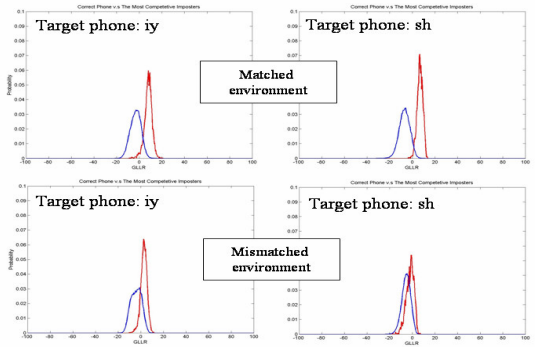

Figure 3: Model separation and acoustic mismatches

Again, we find the GLLR plot of models serves as a good utility to observe model behavior of unseen data by simulating adverse conditions. New compensation algorithms can also be developed to enhance model separation using this utility [1].

3.3. Model separation and training criteria

It is well-known that a set of good models will usually provide a good performance improvement. This improvement can be easily observed using the GLLR utility without running large scale recognition experiments. For example when comparing the conventional maximum likelihood (ML) trained with MCE learned models, we always plot the GLLR statistics before and after MCE training to illustrate the concept of separation enhancement. Here we illustrate this by using a context independent /Vowel/ manner HMMs. In Figure 4, it is clearly shown that the MCE-trained model enhances the separation with its competing models. It is recommended that such GLLR plots are used to compare models trained in various conditions with different optimization criteria.

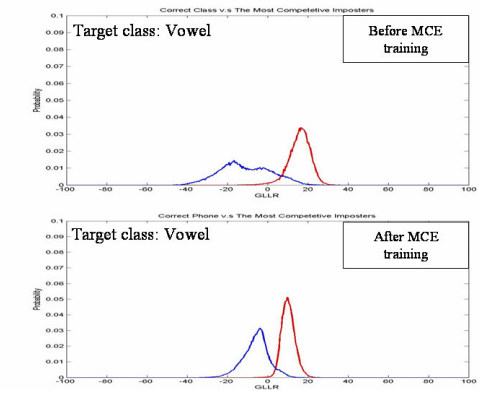

Figure 4: Model separation and training criteria

3.4. Model separation and acoustic resolution

Intuitively a model with a better acoustic resolution will give more separation than models with less detailed description. This can be demonstrated using the GLLR utility to compare context independent (CI) and context dependent (CD) models. Here we used manner attribute models. Our recognition results showed that CD class models reduced the overall class error rate by 18.23% (from 28.91% to 23.64%) when compared with CI class models. In Figure 5 we compared CI /Vowel/ class model with CD /Fricative-Vowel+Stop/ class model. It can be seen that the separation is enhanced with models with a better acoustic resolution, which resulted in a reduction of both Type I and Type II errors.

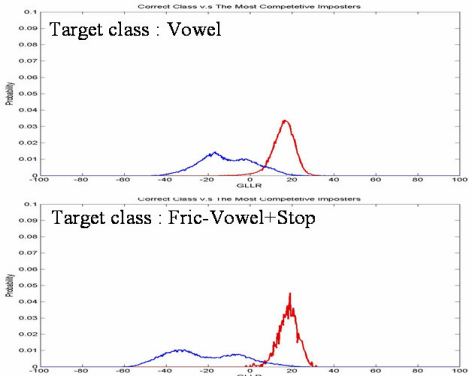

Figure 5: Model separation and acoustic resolution

3.5. Model separation and speech parameter selection

The same GLLR utility can also be used to compare detectors using different speech parameters. It is well known that some

speech parameters are more discriminative in detecting certain speech attributes. A single *voice onset time* (VOT) parameter was shown to give better detection results than those produced with 39 MFCC parameters in differentiating voiced against unvoiced stop sounds [14]. This property can be clearly illustrated by plotting the GLLR histograms to compare the model separation induced by the two sets of detectors using different speech parameters. In Figure 6 (adopted from [15]) for comparing speaker verification parameters, we plot two sets of GLLR histogram plots for one speaker to show that a single pitch parameter gives a smaller overlapping region in the bottom panel than that obtained with 39 MFCC parameters in the top panel, similar to the above VOT case for ASR.

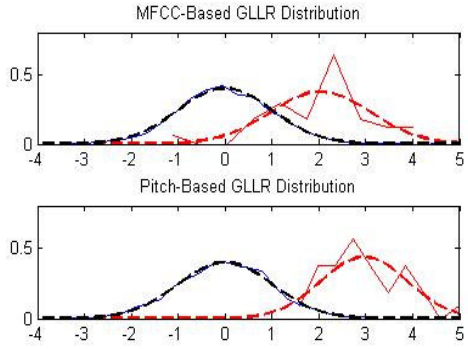

Figure 6: Speaker model separation and MFCC and pitch

Although new speech parameters may not give a significant word error reduction in a complex large vocabulary continuous speech recognition task, the GLLR measure is still a useful tool to evaluate these speech parameters in a well-controlled testing environment in order to demonstrate its utility in discriminating special classes of sounds. We believe that it is critical to develop class-specific speech parameters and fuse them to provide different recognition capabilities. The detection-based ASAT paradigm is an ideal framework to accommodate a large set of diverse speech parameters for speech attribute detection and automatic speech recognition.

4. Summary

The separations between models are closely related to the performance of pattern recognition and verification systems. A one-dimensional GLLR measure can be used to measure the distance between a target and a set of competing models. We found that the two histograms corresponding to the GLLR statistics derived from a collection of target and non-target samples form a GLLR plot that serves as a useful tool to visually analyze the separation between models. We also found that Type I and Type II errors can be clearly displayed on a GLLR plot and two sets of GLLR plots corresponding to two given sets of models can be compared to estimate the discrimination power and the implied recognition or verification errors. We illustrate five examples of the GLLR measure and demonstrate their potential extensions to different applications. We believe the GLLR measure serves as a great tool for developing algorithms based on improved speech models, and new speech parameters without having to conduct large scale, real world ASR and UV experiments.

5. Acknowledgements

Part of this effort was supported under the NSF SGER grant, IIS-03-96848, and the NSF ITR grant, IIS-04-27413. We also thank Chengyuan Ma of Georgia Tech for sharing Figure 6.

6. References

[1] Gauvain, J.-L., and Lamel, L., "Large-Vocabulary Continuous Speech Recognition: Advances and Applications," *Proc. IEEE*, Vol. 88, No. 8, pp. 1181-1200, Aug. 2000.

[2] Sukkar, R. A., and Lee, C.-H., "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. on Speech and Audio Proc.*, Vol. 4, No. 6, pp. 420-429, Nov. 1996.

[3] Huang, C.-S., Wang, H.-C., and Lee, C.-H., "A Study on Model-Based Error Rate Estimation for Automatic Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. 11, pp. 581-589, Nov. 2003.

[4] Juang, B.-H., Chou, W., and Lee, C.-H., "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. 5, pp. 257-265, 1997.

[5] Rahim, M., and Lee, C.-H., "String-Based Minimum Verification Error (SB-MVE) Training for Speech Recognition," *Computer Speech and Language*, Vol. 11, pp. 147-160, 1997.

[6] Lehmann, E. L., *Testing Statistical Hypothesis*, Wiley, New York, 1959.

[7] Katagiri, S., Juang, B.-H. and Lee, C.-H., "Pattern Recognition Using A Generalized Probabilistic Descent Method," *Proc. IEEE*, Vol. 86, No. 11, pp. 2345-2373, Nov. 1998.

[8] Lee, C.-H., "From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition," *Proc. ICSLP2004*, Jeju, South Korea, Oct. 2004.

[9] Lee, C-H. "A Tutorial on Speaker and Speech Verification," *Proc. NORSIG*, Vigso, Denmark, 1998.

[10] Rosenberg, A. E., Delong, J., Lee, C.-H., Juang, B.-H., and Soong, F.K., "The Use of Cohort Normalized Scores for Speaker Recognition," *Proc. ICSLP-92*, pp. 599-602, Banff, Oct. 1992.

[11] Garofolo, J. S. *et al.*, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, February 1993.

[12] Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE,* Vol.77, No.2, pp. 257-286, 1989.

[13] Kirchhoff, K., "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," *Proc. ICSLP98*, Sydney, Australia, 1998.

[14] Niyogi, P. and Ramesh, P., "A Detection Framework for Locating Phonetic Events," *Proc. ICSLP*, Sydney, 1998.

[15] Ma, C. and Lee, C.-H., "Speaker Verification Based on Combining Speaker Parameter Selection and Decisions," *submitted to InterSpeech* 2005.