Towards Structured Approaches to Arbitrary Data Selection and Performance Prediction for Speaker Recognition

Howard Lei

The International Computer Science Institute, Berkeley, CA
hlei@icsi.berkeley.edu
http://www.icsi.berkeley.edu

Abstract. We developed measures relating feature vector distributions to speaker recognition (SR) performances for performance prediction and potential arbitrary data selection for SR. We examined the measures of mutual information, kurtosis, correlation, and measures pertaining to intra- and inter-speaker variability. We applied the measures on feature vectors of phones to determine which measures gave good SR performance prediction of phones standalone and in combination. We found that mutual information had an -83.5% correlation with the Equal Error Rates (EERs) of each phone. Also, Pearson's correlation between the feature vectors of two phones had a -48.6% correlation with the relative EER improvement of the score-level combination of the phones. When implemented in our new data-selection scheme (which does not require a SR system to be run), the measures allowed us to select data with 2.13% overall EER improvement (on SRE08) over data selected via a brute-force approach, at a fifth of the computational costs.

Key words: Text-dependent speaker recognition, mutual information, relevance, redundancy, data selection

1 Introduction

Conversational speaker recognition (SR) involves the task of determining whether a certain speaker spoke in a certain conversation. SR systems have historically relied on GMM speaker models [1], and involves scoring test utterances against target speaker models to determine if the target speaker spoke in the test utterance. The Equal Error Rate (EER), which represents the rate at which false accepts equal false rejects, is a common SR system evaluation standard.

Unit-based text-dependent speaker recognition (UTSR) is the speaker recognition approach where only certain units (i.e. words, phones, syllables) found in speech data are used to construct entire speaker recognition systems [2]. These approaches have been successfully applied in conversational speaker recognition tasks, where the data consists of lengthy conversations between speakers, and the speech is not lexically constrained [2][3]. While discarding much of the speech, the advantages of UTSR for conversational speaker recognition (SR) are threefold: to focus speaker modeling power on more informative regions of speech, to reduce intra-speaker lexical variability, and to reduce the total amounts of data required for faster processing.

The units examined in the past include word N-grams, syllables, phones, and Automatic Language Independent Speech Processing (ALISP) units [4] (which are designed to mimic the phones) and MLP-based phonetic units [5]. Many of the units, such as the words and phones, are used only because their transcripts are readily available via Automatic Speech Recognition, and are incorporated without regard to their actual speaker discriminative abilities. Moreover, there has been no evidence suggesting that words, phones, and/or syllables are ideal sets of units for UTSR. The eventual aim of this work is to allow one to step beyond the use of these units, and to examine the speaker discriminative capabilities of all possible speech segments that can act as units.

This work involves the development of measures as computationally inexpensive ways of determining which units are speaker discriminative based solely on feature vectors of the units. The measures would allow for a quick determination of SR performances of each unit without having to run the SR system, which could take days depending on the units used. For an arbitrary set of units, one task is to compute the measures on the feature vectors of each unit separately. Measures computed in this matter (referred to as relevance measures) would give an indication of the relevance of the unit with respect to the SR task. Measures that have high correlation (in magnitude) with the SR EERs of the units would have good predictive value for SR, and would eventually be good measures for arbitrary data selection.

In UTSR, the units are usually combined at either the feature-level or score-level. To get a good prediction of the effectiveness of unit combination, another task is to compute the measures on pooled features for sets of units, so that a correlation between the measures and the EER achieved via the combination of the set of units is obtained. Measures computed in this manner (referred to as redundancy measures) give an indication of the redundancy of the units amongst one another, whereby units that combine well are less redundant, and vice versa.

Finding effective relevance and redundancy measures will allow for the eventual selection of arbitrary sets of units that produce the best SR performances. Note that the task of data selection is more difficult than the related task of feature selection, in that there are typically many more feature vectors than the number of feature dimensions in a speech utterance, and there are no pre-defined orderings of the feature vectors as opposed to the feature dimensions.

This paper is organized as follows: Section 2 describes the database and our SR system for computing the EERs, section 3 describes the measures, section 4 describes our data-selection scheme, section 5 describes the units used, section 6 describes the experiments and results and provides a brief discussion, and section 7 provides a summary of the current work and describes the applicability of this work to future research in UTSR.

2 Data, Preprocessing, and Speaker Recognition

We used the Switchboard II and Fisher corpora for universal background speaker model training, SRE06 for development, and SRE08 for testing. All corpora consists of telephone conversations between two unfamiliar speakers. A conversation side (roughly 2.5 minutes for non-Fisher and 5 minutes for Fisher) contains speech from one speaker only. 1,060 conversation sides with 128 speakers are used for SRE06, and 1,108 conversation sides with 160 speakers for SRE08. 1,553 background conversation sides are used from Switchboard II and Fisher. Only female English telephone conversation sides are used for this work. There are ∼55,000 total trials for SRE06 with ∼7,000 true speaker trials, and ∼47,000 trials for SRE08 with ∼6,500 true speaker trials. We are provided with force-aligned phone ASR decodings for all conversation sides by SRI, obtained via the DECIPHER recognizer [6].

A 512-mixture GMM-UBM system [1] with MAP adaptation and MFCC features C0-C19 (with 25 ms windows and 10 ms intervals) with deltas is used for computing the EERs of units. The ALIZE implementation is used [7], and the MFCC features are obtained via HTK [8].

3 The Measures

We have implemented various measures for determining the relevance and redundancy of units. For the relevance task, we want to determine how well the measure(s), when computed using the feature vectors of a unit, correlate with the SR performance of the unit. For the redundancy task, we want to determine how well the measure(s), when computed using the feature vectors of a pair of units, correlate with the EER improvement of the MLP-based combination of SR scores of the unit pairs. The measures include mutual information, kurtosis, intra- and inter-speaker variances, Fisher's ratio, and Pearson's correlation.

3.1 Mutual Information as Relevance Measure

Mutual information, which measures the mutual dependence of two variables, has historically been used successfully in the related area of feature selection, such as in [9], [10], and [11]. Typical feature selection algorithms involving mutual information select features with high mutual information with respect to a classification label or class, such that the features are relevant to the classification task. For the case of SR, the classification classes (which are discrete) are the distinct speakers. The mutual information between a continuous vector $\boldsymbol{X}$ and a discrete classification label Y (with distributions $p(\boldsymbol{x})$, $p(y)$, and $p(\boldsymbol{x}, y)$), is given as follows:

$$I(\boldsymbol{X}; Y) = \sum_{y} \int_{\boldsymbol{x}} p(\boldsymbol{x}, y) \log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)} d\boldsymbol{x}. \tag{1}$$

For this particular work, the mutual information is computed between the feature vectors ($\boldsymbol{X}$) constrained by a particular unit, and the speakers (Y). The technique described in [9] for computing the mutual information via approximation of equation 1 is used here. The Parzen windowing technique is used to model the distribution of the continuous-valued feature vectors:

$$\hat{p}(\boldsymbol{X}) = \frac{1}{S}\sum_{i=1}^{S}\delta(\boldsymbol{x}-\boldsymbol{x}_i, h). \tag{2}$$

where $\delta(\cdot)$ is the Parzen window function [12]. A Gaussian window is used, where h represents the standard deviation.

3.2 Kurtosis as Relevance Measure

Kurtosis is a measure of peakiness and/or non-Gaussianity of a random variable. Kurtosis mismatches between training and test conversation sides have been shown to adversely affect speaker recognition performance, and kurtosis feature normalization is an effective way to improve speaker recognition performance [13]. Kurtosis is defined for random variable X as:

$$Kurtosis(X) = \frac{E(x^4)}{E(x^2)^2} - 3 \tag{3}$$

For this work, the kurtosis measure is evaluated on the entire set of feature vectors for each unit.

3.3 Fisher's Ratio, Intra- and Inter-speaker variances as Relevance Measures

Fisher's ratio and intra- and inter-speaker variances all give measures of class-separability, whereby features/data with high Fisher's ratio, high inter-speaker variances, and low intra-speaker variances have high relevance with respect to the classification task. For this work, Fisher's ratio is the ratio of the inter- to intra- speaker variances of the feature vectors of a unit, where we estimated the inter-speaker variance as follows:

$$\sum_{speaker:s}(\boldsymbol{\mu}_s-\boldsymbol{\mu})^T(\boldsymbol{\mu}_s-\boldsymbol{\mu}). \tag{4}$$

and the intra-speaker variance as follows:

$$\sum_{speaker:s}\frac{1}{N_s}\sum_{i\in s}(\boldsymbol{x}_i-\boldsymbol{\mu}_s)^T(\boldsymbol{x}_i-\boldsymbol{\mu}_s). \tag{5}$$

where N_s and $\boldsymbol{\mu}_s$ are the number and average of feature vectors respectively for speaker s, $\boldsymbol{\mu}$ is the overall average of the feature vectors, and $\boldsymbol{x}_i$ is feature vector i.

3.4 Pearson's Correlation as Redundancy Measure

For a pair of units, Pearson's correlation is computed using the average MFCC feature values of each unit for each conversation side. Specifically, for each conversation side, the average values of the MFCC feature vectors for each unit are computed. Pearson's correlation between the averaged values of each unit is computed across all conversation sides. Note that the correlation is computed separately for each dimension of the feature vectors, and an overall correlation is obtained by averaging the correlations of each dimension. Fig. 1 illustrates this computation.

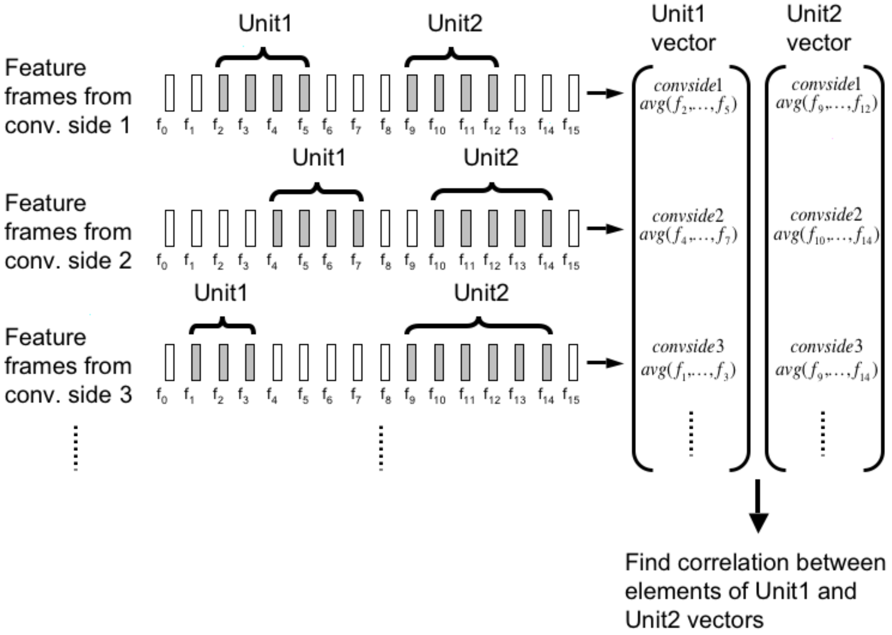
Unit1
Unit2
Unit1 vector
Unit2 vector
Feature frames from conv. side 1
f_0 f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10} f_{11} f_{12} f_{13} f_{14} f_{15}
$convside1$ $avg(f_2,\dots,f_5)$
$convside1$ $avg(f_9,\dots,f_{12})$
Unit1
Unit2
Feature frames from conv. side 2
f_0 f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10} f_{11} f_{12} f_{13} f_{14} f_{15}
$convside2$ $avg(f_4,\dots,f_7)$
$convside2$ $avg(f_{10},\dots,f_{14})$
Unit1
Unit2
Feature frames from conv. side 3
f_0 f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10} f_{11} f_{12} f_{13} f_{14} f_{15}
$convside3$ $avg(f_1,\dots,f_3)$
$convside3$ $avg(f_9,\dots,f_{14})$
Find correlation between elements of Unit1 and Unit2 vectors

Fig. 1. Illustration of the procedure for computing Pearson's correlation as a redundancy measure.

Hence, a Pearson's correlation value is associated with each pair of units. The correlation between this correlation and the relative MLP-based score level combination improvement of the unit pair is obtained to determine how well the measure predicts the redundancy of the unit pair. The relative MLP-based score level combination is determined by the relative score-level combination EER improvement over the average EER of the units standalone.

Note that we've also implemented mutual information as a redundancy measure, but found that Pearson's correlation is more effective.

4 Data Selection Scheme involving the Measures

Our data selection scheme involving the measures is based off of the feature selection approach in [10]. Specifically, given a set of units, the task is to select N units that produce the best SR result in combination. Given the relevance measures for each unit and redundancy measures for unit pairs, our data selection approach is the following: for a given set of pre-selected units P, determine if an additional unit Q should be selected by maximizing the following objective OBJ:

$$OBJ(Q) = Rel(Q) - \alpha \sum_{p \in P} Red(Q, p). \tag{6}$$

where $Rel(Q)$ is the value of the relevance measure for unit Q, $Red(Q, p)$ is the value of the redundancy measure between Q and p, and α is a weight between the relevance and redundancy factors. This objective allows one to select units that have good standalone speaker discriminative power (according to $Rel(Q)$) and are not redundant in their speaker discriminative characteristics with pre-selected units.

5 The Units

The following set of 30 phones represent the units used for this work: /ɑ/, /æ/, /ʌ/, /ɔ/, /ə/, /ɑ^y/, /b/, /d/, /ð/, /ɛ/, /ɝ/, /e/, /f/, /h/, /I/, /i/, /k/, /l/, /m/, /n/, /o/, /p/, PUH, /r/, /s/, /t/, /u/, /v/, /w/, /j/, /z/, where PUH is the vowel in a filled pause, and the remaining phones are denoted by their IPA symbols. These phones are selected from the set of all phones because they occur most frequently in the SRE06 conversation sides.

Phones intuitively represent a good starting point for the evaluation of measures because they span the vast majority of the acoustic space of speech. Each phone represents a small section of the acoustic space, largely separate from the acoustic spaces of other phones. Hence, the use of phones allows the measures to be computed on many different parts of the acoustic space, and the value obtained for a measure using a particular phone would be largely specific to the section of acoustic space represented by the phone. Note that for each conversation side, some phone instances are removed to ensure roughly equal numbers of frames for all phones in the conversation side.

6 Experiments and Results

6.1 Mutual Information as Relevance Measure

Mutual information as a relevance measure is implemented for each of the 30 phones on SRE06 with 128 female speakers, and is our most effective measure. A -0.8352 correlation is obtained between the mutual informations and EERs of the phones. This correlation implies that in general, phones with good SR performance (low EER) also have high mutual information, and that mutual information is an effective measure for SR performance prediction. The phones with the lowest EER and highest mutual information involve the nasals, some consonants, and no vowels: /t/, /k/, /s/, /p/, /f/, /v/, /d/, /ð/, /z/, /b/, /m/, /n/. It is hypothesized that the use of delta features, which capture transitions into and out of the consonants, may have improved the speaker discriminative abilities of the consonants.

The following 6 phones (according to their IPA symbols): /t/, /d/, /b/, /m/, /I/, /ɛ/ resulted in a -0.9969 correlation between mutual information and EER on the SRE06 corpus. The same phones show a -0.9819 correlation on the SRE08 corpus, suggesting that if all speech data are comprised of the 6 phones, a good indication of which phones are speaker discriminative based on their individual mutual information values can be achieved.

6.2 Kurtosis, Fisher's Ratio, Intra- and Inter-speaker Variances as Relevance Measures

Kurtosis, Fisher's ratio, and intra- and inter-speaker variances are also computed on each of the 30 phones, and their values are compared to the EERs of the phones. SRE06 results for the correlations of kurtosis, Fisher's ratio, and intra- and inter-speaker variances for each phone with respect to the EERs are shown in table 1. The result for mutual information is shown alongside these results.

Table 1. Correlations of kurtosis, Fisher's ratio, intra- and inter-speaker variances, and mutual information for each phone with their EERs. Results obtained on SRE06.

Measure	Value
Kurtosis	0.715
Fisher's ratio	0.363
Intra-speaker variance	0.580
Inter-speaker variance	0.539
Mutual information	-0.835

According to table 1, mutual information and kurtosis have the most significant correlations (-0.835 and 0.715 respectively) with the EERs of the 30 phones. Note that the correlation between inter-speaker variance and EER is positive, which is counterintuitive, since the inter-speaker variance should be high for phones with good speaker discriminative ability (and hence low EER). While this is rather strange, past results on Nuisance Attribute Projection (NAP) have suggested that minimizing inter-speaker variance helps SR performances [14][15]. One possible explanation for this is that features with high inter-speaker variance also have high intra-speaker variance in general (this has been shown by examining plots of the feature vectors along the top 2 PCA dimensions for speaker pairs). Nevertheless, these results demonstrate a significance in the correlations between all measures and EER (with the possible exception of Fisher's ratio, which only has a correlation of 0.363). Thus, the measures are useful for SR performance prediction.

6.3 Pearson's Correlation as Redundancy Measure

The approach described in section 3.4 is implemented on the 128 female speakers of SRE06. Correlations are obtained between the feature vectors of all distinct pairs of the 30 phones, along with the relative improvement in the MLP-based score-level combinations of the pairs (two SRE06 splits are created; MLP weights are trained using one split and tested on the other). The latter is obtained by computing the EER improvements of phones in combination over the average of the standalone phone EERs.

The optimal correlation between the correlation of feature vectors and the EER relative improvements of phone pairs is -0.486, which is obtained by considering only C0 and C1 of the MFCC feature vectors without their deltas (a -0.409 correlation is obtained when considering all MFCC coefficients). This result suggests that if the correlation between feature vectors of two phones is high, then the relative improvement of their score-level combination is low, and vice versa. Hence, Pearson's correlation is a suitable measure of unit redundancy.

6.4 Data Selection Investigation and Discussion

We've applied the mutual information relevance measure (our best measure) and Pearson's correlation redundancy measure to the data selection scheme described in section 4. Obtaining the mutual information and Pearson's correlation measures requires roughly a fifth of the computational costs of running the SR system for all phones. We've also used the standalone EERs of the individual phones as the baseline relevance measure. Only C0 and C1 are used for Pearson's correlation measure, which produces the optimal correlation according to section 6.3. All measures (including the standalone EERs) are obtained on SRE06. Two splits of SRE06 are used to train the α parameter from equation 6.

The data selection scheme in section 4 is used to select the top 5 phones for MLP-based score-level combination on SRE08 (with MLP weights trained on SRE06). We've also selected the top 5 phones with the lowest standalone EERs for SRE08, and compared the phone combination EERs obtained via the two approaches. Table 2 shows the EER results on SRE08 for α equal to its optimal value (where both relevance and redundancy are used) and zero (where only relevance is used), along with the phones selected.

According to table 2, selecting the top 5 phones in combination using our data selection approach with mutual information relevance measure and optimal α gives a 13.8% EER on SRE08, which is a 2.13% improvement over selecting the top 5 phones with the best EERs (14.1% EER). Note that even though the improvement is not significant, we've shown that we can select an effective set of units without having to run the actual SR system. Our result also achieves a 4.83% improvement over using mutual information and no redundancy measure

Table 2. MLP score-level combination of top 5 phones selected according to relevance and redundancy measures with optimal α, and standalone EERs. Results obtained on SRE08.

Data selection approach	Relevance measure	Phones selected	EER (%)
Relevance and redundancy	Mutual information	/d/, /h/, /k/, /t/, /v/	13.8
Relevance only	Mutual information	/f/, /k/, /p/, /s/, /t/	14.5
Relevance and redundancy	Standalone EERs	/b/, /k/, /n/, /t/, /z/	13.5
Relevance only	Standalone EERs	/b/, /d/, /k/, /t/, /z/	14.1
Top standalone EERs	–	/b/, /d/, /k/, /t/, /z/	14.1

(14.5% EER), and is within 2.22% of the result using standalone phone EERs as the baseline relevance measure (13.5% EER).

Table 2 also suggests that using only the mutual information relevance measure (with no redundancy measure) for data selection does not improve results over using phones with top standalone EERs (14.5% EER vs. 14.1% EER). The latter can be expected, since the mutual information relevance measure has only an imperfect correlation with EER (-0.835). Note from table 2 that the top 5 phones are all consonants.

According to the results, we have demonstrated that it is possible to select effective units for SR without running the actual SR system, by obtaining relevance and redundancy measures (mutual information and Pearson's correlation in our case) from acoustic feature vectors with good SR performance predictions. Note that using the phone EERs as a baseline relevance measure requires running the SR system, but improves only insignificantly over using mutual information. Our results indicate that taking both relevance and redundancy (as opposed to just relevance) into consideration for SR data selection leads to better unit selection. Interestingly, only the MFCC C0 and C1 coefficients are sufficient for computing our redundancy measure.

7 Conclusion and Future Work

In this work, we've investigated the feasibility of obtaining measures for data selection and performance prediction for unit-based text-dependent speaker recognition. As a starting point, we've used a set of 30 phones as units, and obtained various measures having significant correlations with EERs of the phones. We've shown that it is possible to select a set of units based off of relevance and redundancy measures, which gives equal or better speaker recognition results in combination than the combination of units with best standalone EERs, and does not require a speaker recognition system to be run.

In the future, we will attempt to develop more effective measures, investigate data selection using other types of units, and investigate other types of features. Once we are satisfied with the effectiveness of our measures, we will use the measures to select an arbitrary set of units which would have the globally optimal speaker recognition result in combination for particular types of systems. The arbitrary selection of units would be computationally feasible via the use of our measures.

8 Acknowledgements

The author wishes to thank Andreas Stolcke of SRI for providing speech recognition decodings. This research is funded by NSF grant number 0329258.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.: Speaker Verification using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
2. Sturim, D., Reynolds, D., Dunn, R., Quatieri, T.: Speaker Verification using Text-Constrained Gaussian Mixture Models. In: ICASSP, Vol. 1, pp. 677–680 (2002)
3. Lei, H., Mirghafori, N.: Word-Conditioned Phone N-grams for Speaker Recognition. In: ICASSP, Vol. 4, pp. 253–256 (2007)
4. Hannani, A., Toledano, D., Petrovska-Delacrétaz, D., Montero-Asenjo, A., Hennebert, J.: Using Data-driven and Phonetic Units for Speaker Verification. In: IEEE Odyssey (2006)
5. Gerber, M., Beutler, R., Pfisher, B.: Quasi Text-Independent Speaker-Verification based on Pattern Matching. In: Interspeech, pp. 1993–1996 (2007)
6. Stolcke, A., Bratth, H., Butzberger, J., Franco, H., Rao Gadde, V., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F., Zheng, J.: The SRI March 2000 Hub-5 Conversational Speech Transcription System. In: NIST Speech Transcription Workshop (2000)
7. Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free Toolkit for Speaker Recognition. In: ICASSP, Vol. 1, pp. 737–740 (2005)
8. HMM Toolkit (HTK), `http://htk.eng.cam.ac.uk`
9. Kwak, N., Choi, C.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002)
10. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005)
11. Ellis, D., Bilmes, J.: Using Mutual Information to Design Feature Combinations. In: ICSLP (2000)
12. Parzen, E.: On Estimation of a Probability Density Function and Mode. Annals of Math. Statistics 33 (1962)
13. Xie, Y., Dai, B., Yao, Z., Liu, M.: Kurtosis Normalization in Feature Space for Robust Speaker Verification. In: ICASSP, Vol 1 (2006)
14. Vogt, R., Kajarekar, S., Sridharan, S.: Discriminant NAP for SVM Speaker Recognition. In: IEEE Odyssey (2008)
15. Lei, H.: NAP, WCCN, a New Linear Kernel, and Keyword Weighting for the HMM Supervector Speaker Recognition System. Technical report, International Computer Sciences Institute (2008)