Importance of Nasality Measures for Speaker Recognition Data Selection and Performance Prediction

Howard Lei[1], *Eduardo Lopez-Gonzalo*[1,2]

[1]The International Computer Science Institute, Berkeley, CA
[2]Signal Processing Applications Group, Universidad Politécnica de Madrid, Madrid, Spain
hlei@icsi.berkeley.edu, eduardo@icsi.berkeley.edu

Abstract

We improve upon our measures relating feature vector distributions to speaker recognition (SR) performances for performance prediction and potential arbitrary data selection for SR, as described in [1]. In particular, we examine the means and variances of 11 features pertaining to nasality (each of which is denoted as a measure), computing them on feature vectors of phones to determine which measures give good SR performance prediction of phones. We've found that the combination of nasality measures give a 0.917 correlation with the Equal Error Rates (EERs) of phones on SRE08, exceeding the correlation of our previous best measure (mutual information) by 12.7%. When implemented in our data-selection scheme (which does not require a SR system to be run), the nasality measures allow us to select data with combined EER better than data selected via running a SR system in certain cases, at a fortieth of the computational costs. The nasality measures also require a tenth of the computational costs to compute compared to our previous best measure.

Index Terms: Text-dependent speaker recognition, data selection, nasality measures, relevance, redundancy

1. Introduction

Unit-based text-dependent speaker recognition (UTSR) is the speaker recognition approach where only certain speech 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 three-fold: 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.

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. In [1], we showed the promise of our approach by demonstrating that measures such as mutual information, kurtosis, and Pearson's correlation led to effective data selection. In this work, we improve upon our previous work by examining measures pertaining to nasality and their effectiveness in data selection.

This paper is organized as follows: Section 2 describes the database and our SR system for computing the EERs, section 3 describes the nasality 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 8 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, and

1,180 conversation sides with 128, and 160 speakers are used for SRE06, and SRE08 respectively. 1,553 background conversation sides are used from Switchboard II and Fisher. One conversation side is used to train each target speaker model, and 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. In addition, we created two splits of SRE06 (SRE06s1 and SRE06s2) for development purposes. There are ∼65 speakers, ∼530 conversation sides, ∼15,000 trials, and ∼3,400 true speaker trials in each split. 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 [7] 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 [8], and the MFCC features are obtained via HTK [9].

3. Nasality features as relevance measures

Previous work suggests that nasal regions of speech are an effective speaker cue, because the nasal cavity is both speaker specific, and fixed in the sense that one cannot change its volume or shape [10]. Various acoustic features have been proposed for detecting nasality. Glass used six features for detecting nasalized vowels in American English [11]. Pruthi extended Glass's work and selected a set of nine knowledge-based features for classifying vowel segments into oral and nasal categories automatically [12].

Our goal, however, is to determine if the nasality features would allow us to identify which speech units have good speaker discriminative power. The fact that the features have been used to detect nasalization in vowels would possibly allow the features to better determine which speech units hold greater speaker discriminative power, since nasals themselves hold great speaker discriminative power [10]. The means and variances of each nasality feature, computed over all data constrained by a speech unit, are used as relevance measures for that unit.

3.1. Description of nasality features

All nasality features described below are computed using 25 ms windows with 10 ms shifts. A total of 11 nasality features are implemented.

std01k: The standard deviation of frequency around the center of mass of the frequency region below 1000Hz. Standard deviation is calculated using the spectral amplitudes 500 Hz on each side of the center of mass, but constrained to within 0 and 1000 Hz [11].

ctm01k: The center of mass of the short-term log magnitude squared (dB) spectrum amplitude in the frequency band between 0 and 1000 Hz. It is computed using a trapezoidal window with flatness between 100-900Hz.

a1h1max800: The difference, measured in the log magnitude squared spectrum, between the amplitude of the first formant (A1) and the first harmonic (H1) [12]. A1 is estimated using the amplitude of the maximum value in the band between 0 and 800 Hz. H1 is obtained using the amplitude of the peak closest to 0Hz which had a height greater than 10dB and a width greater than 80Hz.

a1max800: The amplitude of the first formant (A1) relative to the total spectral energy between 400 Hz and 800 Hz.

tef1: The teager energy operator for detection of hypernasality [13]. It finds the correlation between the teager energy profiles of narrow bandpass-filtered speech and wide bandpass-filtered speech centered around the first formant.

c0: The 0th cepstral coefficient representing the energy of the spectrum. Our intuition is that this feature would be smaller on average for nasals because nasals appear to be softer in amplitude in general.

frat: The ratio of the spectral energies between 300 to 700 Hz and between 2,500 to 3,400 Hz. We observed the ratio to be higher on average for nasals.

Four additional features are extracted based on the detection of possible poles below and above the first formant. These poles are computed using a smoothed version of the FFT spectra. Denote *p0* and *fp0* as the amplitude and frequency of the pole below the first formant, *p1* and *fp1* as the amplitude and frequency of the pole above the first formant, and *a1* and *f1* as the amplitude and frequency of the first format. The features are *a1-p0*, *a1-p1*, *f1-fp0* and *fp1-f1*.

3.2. Pearson's correlation as redundancy measure

We've used the same redundancy measure described in our previous work [1]. For a pair of units, Pearson's correlation is computed using the average MFCC feature values of each unit for each conversation side. For each conversation side, the average values of the MFCC feature vectors for each unit are computed. A correlation across all conversation sides is then computed between the averaged MFCC feature values of unit 1 and unit2. 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.

Each pair of units is thus associated with a correlation (denote by *corrA*). We have shown in [1] that there is a -0.486 correlation between the unit pair's relative improvement in score-level combination, and its *corrA* value, indicating that *corrA* is a valid indicator of the unit pair's redundancy.

4. Data selection scheme involving the measures

Our data selection scheme is based off of the feature selection approach in [14]. 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:

$$OBJ(Q) = Rel(Q) - \alpha \sum_{p \in P} Red(Q, p). \qquad (1)$$

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 character-

istics 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/, /ɑʸ/, /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. Hence, the use of phones allows the measures to be computed on many different segments 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. Nasality features as relevance measure

As discussed in 3, the mean and variance of each of the 11 nasality features constrained by a unit are used as relevance measures for that unit. We also computed the EER of each unit by running our speaker recognition system using data constrained by that unit, such that each unit is associated with 11 nasality means and variances (22 relevance measures total), and 1 EER. For the 30 units, we've computed the correlations between the EERs and each relevance measure, obtaining a total of 22 correlations. A greater correlation in magnitude indicates a greater ability of the relevance measure in predicting the EER. Table **??** shows the correlations of each relevance measure on SRE08.

The correlation of our mutual information measure (our previous best in terms of correlation with phone EERs) described in [1], is also shown. For each phone, the mutual information is computed between the feature vectors (MFCC C0-C19 + delta) and speakers. It represents the total entropy of the feature vectors minus the entropy of the feature vectors given the speaker.

According to table 1, the *a1h1max800* mean (0.807 correlation) and *tef1* variance (-0.757 correlation) are nasality measures able to most strongly predict the EER (these correlations are significant at the 1% level). However, the individual measures themselves do not outperform mutual information (-0.814 correlation).

The measures are combined via linear regression and stronger correlations between the EER and combined measures are obtained. Leave-one-out (LOO) selection is used to select the most useful set of measures. LOO selection selects the best set of measures with respect to their correlations on SRE06s2, with regression weights trained on SRE06s1. Table 2 shows the correlation on SRE06s2 for 10 iterations of LOO selection.

Interestingly, the measures that perform the best individually (*a1h1max800* Mean and *tef1* variance) are amongst the first to be dropped via LOO selection. Also, the top 17 measures produce the best correlation (0.947 at iteration 5). We kept the top 17 measures, and trained a linear regression model on SRE06. Applying the model on SRE08, we obtain a correlation of 0.917 between the combination of the 17 nasality measures and EER. This is a 12.7% improvement over mutual information on SRE08. Repeating the above procedure while incorporating the mutual information measure, we obtain a 0.912 correlation on SRE08. Table 3 summarizes these results.

Measure	Mean or Var	Correlation
a1max800	Mean	-0.316
a1max800	Var	-0.465
a1h1max800	Mean	0.807
a1h1max800	Var	0.699
c0	Mean	0.252
c0	Var	0.640
ctm01k	Mean	0.471
ctm01k	Var	-0.502
frat	Mean	0.394
frat	Var	0.340
std01k	Mean	-0.041
std01k	Var	-0.510
tef1	Mean	0.197
tef1	Var	-0.757
a1-p0	Mean	0.373
a1-p0	Var	0.486
a1-p1	Mean	0.086
a1-p1	Var	-0.182
f1-fp0	Mean	0.067
f1-fp0	Var	0.055
fp1-f1	Mean	-0.238
fp1-f1	Var	0.344
Mutual Information	–	-0.814

Table 1: Correlations of the means and variances of each nasality feature with the EERs of each phone. Results obtained on SRE08.

Note that the correlation with nasality and mutual information measures is roughly equivalent to the correlation with nasality measures alone, indicating that mutual information does not contribute to correlation improvements.

7. Data selection with nasality measures

We've applied the combined 17 nasality measures (NAS), the mutual information measure (MI), and Pearson's correlation redundancy measure to the data selection scheme described in section 4. Computing the nasality measures requires a tenth of the computational costs of computing the mutual information measure on all phones. Also, computing the nasality measures and Pearson's correlation measures requires roughly a fortieth of the computational costs of running the SR system for all phones.

The standalone EERs of the individual phones are used as the baseline relevance measure. Two splits of SRE06 are used to train the α parameter from equation 1, using measures obtained on SRE06. The data selection scheme in section 4 is used to select the top 5 and 10 phones for MLP-based score-level combination on SRE08 (with MLP weights trained on SRE06). Table 4 shows the EER results on SRE08 for α equal to its optimal value for the measure and number of phones used, along with the phones selected. Note that nasality measure combination is performed on SRE08 with weights trained on SRE06, and the standalone EERs and mutual information are obtained on SRE08.

According to table 4, selecting the top 10 phones in combination using our data selection approach with the combination

Iteration	Nasality feature eliminated	Correlation
1	*c0* Mean	0.918
2	*a1h1max800* Mean	0.924
3	*tef1* Var	0.938
4	*a1h1max800* Var	0.945
5	*a1-p0* Mean	0.947
6	*f1-fp0* Mean	0.947
7	*fp1-f1* Var	0.946
8	*a1max800* Mean	0.946
9	*frat* Mean	0.946
10	*a1-p0* Var	0.945

Table 2: 10 iterations of leave-one-out selection for SRE04s2, with linear regression weights trained on SRE04.

Measure(s)	Correlation
Nasality only	0.917
Mutual information only	-0.814
Nasality + mutual information	0.912

Table 3: Results on SRE08 with and without mutual information measure.

of nasality measures as the relevance measure gives a 11.5% EER on SRE08, which is a 4.96% improvement over using the EERs (12.1% EER) and 1.71% improvement over using mutual information (11.7%) as relevance measures. Note that even though the 1.71% improvement is not significant, is it obtained at a tenth of the computational cost. In addition, we've shown that using the nasality measures, it is possible to select a set of units that perform better than the units selected via running the SR system, at a fortieth of the computational cost.

Even though the standalone EERs (13.5% EER) perform better than the nasality measures (14.5% EER) and mutual information (14.8%) if only 5 phones are to be selected, the computational cost savings imply that using the measures may still be preferred. Nevertheless, we've demonstrated that it is possible to select effective units for SR without running the actual SR system at a very small fraction of the computational cost, and the units selected can perform better in combination that units selected via running the SR system.

8. Conclusion and future Work

In this work, we've investigated the feasibility of using nasality measures for data selection and performance prediction for unit-based text-dependent speaker recognition. Using a set of 30 phones as units, we showed that means and variances of nasality features in combination have significant correlations with phone EERs. We've shown that units selected using the nasality measures as relevance measure can give better speaker recognition results in combination than the combination of units with standalone EERs as relevance measure, and does not require a speaker recognition system to be run.

In the future, we will attempt to develop more effective measures and investigate data selection using other types of units. 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.

Relevance measure	Num phones	Phones selected	EER (%)
MI	5	/ɑ/, /d/, /h/, /k/, /v/	14.8
NAS	5	/b/, /ð/, PUH, /s/, /z/	14.5
EER	5	/b/, /k/, /n/, /t/, /z/	13.5
MI	10	/ɑ/, /d/, /h/, /I/, /k/, /m/, PUH, /s/, /v/, /w/	11.7
NAS	10	/ɑ/, /ɑ^y/, /b/, /d/, /ð/, /m/, /p/, PUH, /s/, /z/	11.5
EER	10	/ɑ/, /æ/, /ɑ^y/, /b/, /d/, /ð/, /k/, /n/, /t/, /z/	12.1

Table 4: MLP score-level combination of top 5 and 10 phones selected according to relevance and redundancy measures with optimal α. Results obtained on SRE08.

9. Acknowledgements

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

10. References

[1] Lei, H.,"Towards Structured Approaches to Arbitrary Data Selection and Performance Prediction for Speaker Recognition", accepted to 3rd International Biometrics Conference, 2009.

[2] Sturim, D., Reynolds, D., Dunn, R. and Quatieri, T.,"Speaker Verification using Text-Constrained Gaussian Mixture Models", in Proc. of ICASSP, 2002.

[3] Lei, H. and Mirghafori, N.,"Word-Conditioned Phone N-grams for Speaker Recognition", in Proc. of ICASSP, 2007.

[4] Hannani, A., Toledano, D., Petrovska-Delacrétaz, D., Montero-Asenjo, A. and Hennebert, J.,"Using Data-driven and Phonetic Units for Speaker Verification", in Proc. of IEEE Odyssey, 2006.

[5] Gerber, M., Beutler, R. and Pfisher, B.,"Quasi Text-Independent Speaker-Verification based on Pattern Matching", in Proc. of Interspeech, 2007.

[6] Stolcke, A., Bratth, H., Butzberger, J., Franco, H., Rao Gadde, V., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F. and Zheng, J.,"The SRI March 2000 Hub-5 Conversational Speech Transcription System", in NIST Speech Transcription Workshop, 2000.

[7] Reynolds, D.A., Quatieri, T.F. and Dunn, R.,"Speaker Verification using Adapted Gaussian Mixture Models", in Digital Signal Processing, pp 19–41, 2000.

[8] Bonastre, J.F., Wils, F., Meignier, S.,"ALIZE, a free Toolkit for Speaker Recognition", in Proc. of ICASSP, 2005.

[9] HMM Toolkit (HTK), http://htk.eng.cam.ac.uk

[10] Amino, K., Sugawara, T. and Arai, T.,"Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties", in Acoustic Science and Technology, 27(4), 2006.

[11] Glass, J.R., Zue, V.W.,"Detection of nasalized vowels in American English", in Proc. of ICASSP, 1985.

[12] Pruthi, T and Espy-Wilson, C. Y.,"Acoustic parameters for the automatic detection of vowel nasalization", in Proc. of Interspeech, 2007.

[13] Cairns, D.A., Hansen, J.H. and Kaiser, J.F.,"Recent advances in hypernasal speech detection using the nonlinear teager energy operator", in Proc. of ICSLP, pp. 780–783, 1996.

[14] Peng, H., Long, F. and Ding, C.,"Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.