Data Selection with Kurtosis and Nasality features for Speaker Recognition

Howard Lei, Nikki Mirghafori

The International Computer Science Institute, Berkeley, CA

hlei@icsi.berkeley.edu, nikki@icsi.berkeley.edu

Abstract

We propose new data selection approaches based on speaker discriminability features, including kurtosis and a set of nasality features which exploit spectral properties of nasal speech sounds. Data selected based on the speaker discriminability features are used to implement end-to-end speaker recognition systems, which produce significant improvements when combined with the baseline system (which uses the speech-only data regions determined by a speech/non-speech detector), where the optimal combination of systems produces roughly a 24% improvement over the baseline. Results suggest that focusing the modeling power on data regions selected via the kurtosis and nasality speaker discriminability features, part of which are often discarded in the speech/non-speech detection process, can improvement speaker recognition.

Index Terms: speaker recognition, kurtosis, nasality features, data selection

1. Introduction

Traditionally, data selection for speaker recognition has either been done using a speech/non-speech detector and keeping the speech regions, or selecting data based on various speech units (i.e. lexical units, such as words, phones, and syllables) [1]. In the former approach, the signal energy is typically used to determine regions of higher energy signal (likely to be speech) versus lower energy signal (likely to be silence). However, this approach may be problematic when confronted with higher-energy, non-speech regions, such as in highly noisy data. In the latter approach, only data corresponding to certain speech units are used to construct end-to-end speaker recognition systems. However, the units that are used usually depend on the existence of an automatic speech recognizer, which may not be available.

In this work, we use a noisy data set to investigate an alternative approach to data selection, which lies along the contiuum spanning the energy-based approaches, and the lexical-based approaches. Instead of using energy, our approach is based on the use of various speaker discriminability features, which have been determined to have good predictive power for the speaker discriminability of lexical regions of speech [2][3][4]. The features include kurtosis and several nasality features, which have been determined to have strong influences on speaker recognition accuracy [2] [3]. These features allows us to select speech data determined to be speaker discriminative.

The features are applied in lieu of speech/non-speech detection, and the data selected via these features can be used to construct end-to-end speaker recognition systems. The advantages of using these features over the strict application of a speech/non-speech detector is that the features select for speaker discriminative regions of utterance data, as opposed to selecting merely the speech-only regions. We presume that in certain contexts, the non-speech regions would have speaker discriminative power, especially in cases where certain speakers are associated with certain environmental conditions. Data selection based on the speaker discriminability features also have potential advantages compared to selection via speech units, in that the data selected are not constrained merely by the units themselves, which may or may not have high speaker discriminative power.

This paper is organized as follows: Section 2 describes the database, section 3 describes the kurtosis and nasality features, section 4 describes our data-selection scheme, section 5 describes the speaker recognition system, section 6 describes the experiments and results and provides a brief discussion, and section 7 provides a summary of the current work and discusses potential future work.

2. Data, preprocessing, and speaker recognition

We used the ROSSI database [5], which contains various types of channel and environmental noise typically with 10 dB SNR per utterance, for our experiments. The portions of the ROSSI database we used consist of utterances of roughly 50 seconds of monologue landline and cellular phone speech, recorded in various noisy environmental conditions. The breakdown of the types of utterances used for UBM training, speaker model training, and testing are shown in table 1.

A total of approximately 100 distinct speakers are used for speaker model training, and 200 speakers are used for testing. 100 of the 200 speakers are used for training, and the remaining 100 do not exist amongst the training speakers. A total of approximately 450,000 speaker recognition trials are used, with 2,500 true speaker trials, and 447,500 impostor trials.

3. The kurtosis and nasality features

Our prior work on speaker discriminability features has indicated that kurtosis and nasality are effective for detecting speaker discriminative regions of speech [4][3][6]. The speaker discriminative power of the kurtosis and nasality features are determined by computing them over regions of 30 phones, and computing the correlation of the feature values to the Equal Error Rates (EERs) obtained from speaker recognition systems implemented using data constrained by each of the phones. Our previous results indicate that the kurtosis feature has a 0.7 correlation with the EERs, while a linear regression of the nasality features produces a 0.9 correlation with the EERs. These results are obtained on 1,060 female conversation sides of the SRE06 database.

Development			
Environment	Channel	# of utterances	# of hours
Office	Landline	200	2.8
Office	Cellular	50	0.7
Public place	Cellular	50	0.7
Vehicle	Cellular	50	0.7
Roadside	Cellular	50	0.7
Total	–	400	5.6
Speaker model training			
Environment	Channel	# of utterances	# of hours
Office	Landline	100	1.4
Office	Cellular	100	1.4
Public place	Cellular	100	1.4
Vehicle	Cellular	100	1.4
Roadside	Cellular	100	1.4
Total	–	500	7
Testing			
Environment	Channel	# of utterances	# of hours
Office	Landline	200	2.8
Office	Cellular	200	2.8
Public place	Cellular	200	2.8
Vehicle	Cellular	200	2.8
Roadside	Cellular	50	0.7
Total	–	850	11.9

Table 1: Description of channel and environment information of utterances in the ROSSI database.

3.1. Kurtosis feature

Kurtosis is a measure of peakiness and/or non-Gaussianity of a random variable. Kurtosis feature normalization is an effective way to improve speaker recognition performance, such that a low kurtosis value (close to 0) is desired [2]. Kurtosis is defined for random variable X as:

$$Kurtosis(X) = \frac{E(x^4)}{E(x^2)^2} - 3 \quad (1)$$

In this work, kurtosis values are computed on MFCC feature vectors using windows of 30 feature frames, with 5-frame shifts. Hence, given that MFCC feature vectors are computed every 10 ms, kurtosis values are computed using feature vectors that span 300 ms, shifting every 50 ms. Kurtosis is computed for each feature dimension separately, and averaged across all feature dimensions to obtain the final kurtosis feature value. Note that selecting data based on kurtosis is similar to Gaussian-normalization of data. However, kurtosis data selection avoids the risk of data distortion through Gaussian normalization by selecting data that's already normalized.

3.2. Nasality features

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 [7]. Various acoustic features have been proposed for detecting nasality. Glass used six features for detecting nasalized vowels in American English [8]. Pruthi extended Glass's work and selected a set of nine knowledge-based features for classifying vowel segments into oral and nasal categories automatically [9].

For this work, we've implemented 5 nasality features that

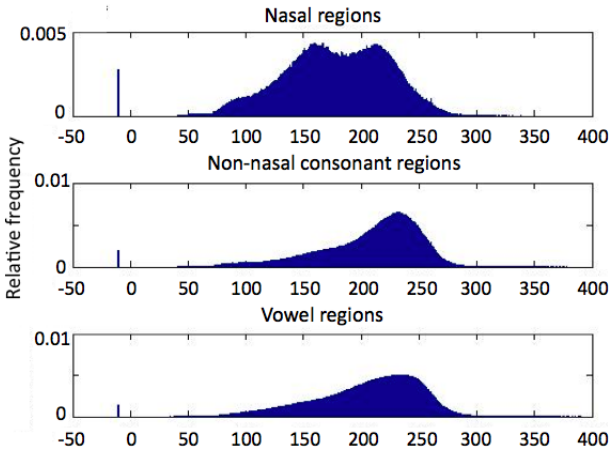

Figure 1: *Distribution of std01k nasality feature for nasal, vowel, and non-nasal consonant regions for 1,060 female speakers in SRE06 database.*

we've determined to be effective for purposes of data selection. All nasality features are computed using 25 ms windows with 10 ms shifts. A description of each are given below:

a1h1max800: The difference, measured in the log magnitude squared spectrum, between the amplitude of the first formant (A1) and the first harmonic (H1) [9]. This feature is found to be lower for nasal regions of speech.

a1max800: The amplitude of the first formant (A1) relative to the total spectral energy between 400 Hz and 800 Hz. This feature is found to be higher for nasal regions of speech.

frat: The ratio of the spectral energies between 300 to 700 Hz and between 2,500 to 3,400 Hz. This feature is found to be higher for nasal regions of speech.

std01k: The standard deviation of frequency around the center of mass of the frequency region below 1000Hz [8]. This feature is found to be lower for nasal regions of speech.

ctm01k: The center of mass of the short-term log magnitude squared (dB) spectrum amplitude in the frequency band between 0 and 1000 Hz. This feature is found to be lower for nasal regions of speech.

As an example of how the nasality features are distributed over speech regions, figure 1 show the distribution of the *std01k* nasality feature when computed over nasal, vowel, and non-nasal consonant regions in 1,060 female SRE06 conversation sides. According to the figure, the *std01k* nasality feature is lower overall respectively when computed across nasal regions, as opposed to the other regions.

4. Selecting data with discriminability features

To select speech data using our features, we set thresholds on the kurtosis and nasality features, and select utterance regions with kurtosis or nasality feature values above or below the thresholds. We know (according to past work) that speech regions with lower kurtosis values, lower *a1h1max800* values, higher *a1max800*, higher *frat*, lower *std01k* and lower *ctm01k* values hold higher speaker discriminative ability. Hence, we

set thresholds such that speech regions with *a1max800* and *frat* values lying above certain thresholds, and with other values below certain thresholds are selected. Lastly, we've attempted to select only the non-speech regions as determined by the Shout speech/non-speech detector [10] to see if they are speaker discriminative in any way.

5. Speaker recognition system

For this work, we've used a 128-mixture GMM-UBM system [11] with MAP adaptation and MFCC features C0-C12 (with 25 ms windows and 10 ms intervals) with deltas and double deltas, and mean and variance normalization. The simplified factor analysis approach is used. The ALIZE implementation is used for GMM model and factor analysis training and testing [12], and the MFCC features are obtained via HTK [13]. We've also used the Shout detector to select the speech regions. The ROSSI development utterances are used for UBM and factor analysis matrix training.

6. Experiments and results

We've applied our data selection scheme using the kurtosis and nasality features on all trials of the ROSSI database, using the aforementioned speaker recognition system. We've first attempted to perform data selection using the kurtosis feature, and tested various lower and upper thresholds to determine the optimal amount of data to keep. The thresholds are represented as percentiles (i.e. a lower threshold of 0.1 indicates that data with the lowest 10 percent of kurtosis values are discarded; an upper threshold of 0.9 indicates that data with the highest 10 percent of kurtosis values are discarded). Results are shown in table 2. For the results with the Shout+Kurtosis feature, the intersection of the data selected via kurtosis and data selected via the Shout speech/non-speech detector is used. For the results with the ALL feature, all utterance data is used.

Feature	Lower threshold	Upper threshold	EER (%)
Kurtosis	0.00	0.70	10.0
Kurtosis	0.00	0.85	9.8
Kurtosis	0.00	0.95	9.9
Kurtosis	0.15	0.85	10.4
Kurtosis	0.15	1.00	10.3
Shout+Kurtosis	0.00	0.85	9.8
ALL	–	–	9.9

Table 2: Results for kurtosis data selection using various thresholds.

Results indicate that using lower and upper thresholds of 0.00 and 0.85 respectively produce the lowest EER (9.8%). This indicates that data with kurtosis values belonging to the lowest 85th percentile should be selected for speaker recognition, which agrees with the fact that speech regions with lower kurtosis values for its feature vectors are more speaker discriminative. We obtain a 4.7% relative EER improvement by taking data selected with kurtosis values in the lowest 85th percentile, versus data selected with kurtosis values in the highest 85th percentile (10.3% EER). Note that while data selection using the Shout+Kurtosis measure also produces a 9.8% EER with the same thresholds, it requires the use of a speech/non-speech detector. Kurtosis-based data selection also leads to a slightly lower EER (though perhaps insignificant) than simply using all utterance data (9.8% vs 9.9% EER).

Based on these results, data selection using the nasality features is also performed by selecting utterance regions with nasality feature values in either the lowest or highest 85th percentile, depending on whether lower or higher nasality feature values indicate greater speaker discriminability.

Table 3 shows the percentage of total utterance data (from all development utterances) selected by both Shout *and* the features, the percentage of total utterance data selected uniquely by Shout and *not* by the features (i.e. regions selected by both were excluded), and the percentage of total utterance data selected uniquely by the features and *not* by Shout. Denote the kurtosis feature as *kurt*.

Feature	% data selected by **Shout** and **feature**	% data sel. by **Shout** and not by feature	% data sel. by **feature** and not by Shout
kurt	61.6	13.4	17.8
a1h1max800	61.6	13.4	23.4
a1max800	60.2	14.8	24.8
frat	67.4	7.6	17.7
std01k	68.3	6.8	16.8
ctm01k	64.4	10.6	20.6

Table 3: Percentage of overlap of data selected via Shout, and via each of the speaker discriminability features, with respect to total utterance data.

Results show that while there is a high degree of correspondence between the data selected using each of the features and using Shout (for each feature, over 60% of all utterance data are selected by the feature *and* by Shout), there are significant regions of data selected by either the features or by Shout, but not both. Moreover, for each feature, the percentage of data selected by the feature and *not* by Shout is greater than the percentage of data selected by Shout and *not* by the feature. This suggests that non-speech regions are more significantly included in the data selected by the features than the speech regions are included in the data selected by Shout. The data selected using the features may thus be complementary to the data selected using Shout.

Each data selection technique (i.e. using kurtosis or nasality features) is used to implement a separate speaker recognition system. EER results are obtained standalone and in combination using an MLP with 2 hidden nodes and 1 hidden layer, implemented using Lnknet [14]. The EERs represent averaged EER values over 100 splits amongst the trials, where each split contains training and testing sub-splits. For each of the 100 splits, MLP weights are trained using the training sub-split, and EERs for each split are obtained by applying the MLP weights on the testing sub-split. This is done even if there is only one system used, so that the standalone results can be consistent with the combination results. The set of all results are shown in table 4, where systems with feature-based data selection are denoted by its feature, the baseline system (which uses the utterance regions determined to be speech according to Shout) is denoted as *base*, and a system that uses only the silence regions obtained using Shout is denoted as *sil*. Note that for each system, factor analysis is applied only in cases where it significantly improves its standalone EER.

Our results indicate that the baseline GMM-UBM system (with factor analysis) using the Shout speech/non-speech detec-

System	EER (%)
base	9.1
sil	32.2
kurt	9.7
a1h1max800	15.2
a1max800	15.0
frat	14.9
base+sil	9.5
base+kurt	**8.0**
a1h1max800+a1max800+frat	7.9
base+kurt+sil+a1h1max800+ a1max800+frat+std01k+ctm01k	7.2
base+kurt+ a1h1max800+a1max800+frat	**6.9**

Table 4: Results for all systems standalone and in combination

tor is the best standalone system, with an EER of 9.1%. The standalone system involving data selection via the kurtosis feature (9.7% EER) is 5.8% worse than the baseline in terms of EER. However, when combined with the baseline system, the kurtosis system produces a 12.5% relative improvement over the baseline system standalone (8.0% EER vs. 9.1% EER).

While the individual nasality features do not perform as well as the baseline system, they perform effectively in combination. Combining systems based on the *a1h1max800*, *a1max800* and *frat* nasality features produces a 7.9% EER, a 14.0% relative improvement over the baseline. We've found that when in combination with the baseline, kurtosis, and silence-based systems, the nasality features contribute to a 7.2% EER, regardless of whether only the *a1h1max800*, *a1max800* and *frat* features are used, or whether all five nasality features are used. Our best result is obtained by combining the baseline system with systems with data selected using the kurtosis, *a1h1max800*, *a1max800* and *frat* features, producing an EER of 6.9%. This represents a 23.9% relative improvement over the baseline standalone.

The system that uses the silence-only regions from the Shout speech/non-speech detector performs surprisingly well, giving a 32.2% EER. This perhaps indicates that there are factors outside of speech (i.e. channel and acoustic environments linked to different speakers) for utterances in the ROSSI database that contribute to speaker recognition accuracy. Overall, the results show that there's useful complementary information in the data selected via the features we used, especially as there are significant differences in the utterance regions that the features selected (refer to table 3).

While past data selection approaches have traditionally been based on selecting only the speech regions, we demonstrate the importance of non-speech regions for speaker recognition in noisy data. The features we implemented, which were able to select utterance regions determined to be speaker discriminative according to past work, allow us to employ a new data selection technique that potentially involves the non-speech regions. Speaker recognition systems implemented based on these regions are complementary to the baseline system using only the speech regions.

7. Conclusion and future work

In this work, we propose new data selection approaches based on speaker discriminability features for noisy speech data. The approaches lie along the contiuum spanning the energy-based approaches, and the lexical-based approaches. The speaker discriminability features include kurtosis, and a set of five nasality features. Data regions where its kurtosis or nasality feature passes certain thresholds are retained for end-to-end speaker recognition system implementation, and the thresholds are determined by the distributions of the features. Combining the baseline system (which uses data selected by a speech/non-speech detector) with the systems using data selected from the speaker discriminability features produces significant improvements over the baseline system standalone. Results suggest that focusing the modeling power on data regions selected via the kurtosis and nasality speaker discriminability features, part of which are often discarded in the speech/non-speech detection process, can improvement speaker recognition. Future work can examine additional approaches to selecting data - such as random data selection - and applying our techniques on standard, larger datasets such as the NIST SRE's.

8. Acknowledgements

This work is sponsored by Air Force Research Laboratory under contract FA8750-10-C-0214. The authors would like to thank Eduardo Lopez-Gonzalo for his work on the nasality features.

9. References

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

[2] Xie, Y., Dai, B., Yao, Z., and Liu, M.,"Kurtosis Normalization in Feature Space for Robust Speaker Verification", in Proc. of ICASSP, 2006.

[3] Lei, H., Lopez-Gonzalo, E., "Importance of Nasality Measures for Speaker Recognition Data Selection and Performance Prediction", in Proc. of Interspeech, 2009.

[4] Lei, H., "Structured Approaches to Data Selection for Speaker Recognition", Ph.D Thesis, UC Berkeley, 2010.

[5] Battles, B. and Lawson, A. "NoTel: A Large, Naturally Noisy, Multi- Device Telephony Database For Speech And Speaker Recognition", Under Review.

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

[7] 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.

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

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

[10] Huijbregts, M., "Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled", Ph. D Thesis, University of Twente, 2008

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

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

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

[14] Lippmann, R.P., Kukolich, L.C., Singer, E., "LNKnet: Neural Network, Machine Learning, and Statistical Software for Pattern Classification", in Lincoln Laboratory Journal, Vol. 6, pp 249-268, 1993.