DEVELOPING A SPEAKER IDENTIFICATION SYSTEM FOR THE DARPA RATS PROJECT

Oldřich Plchot [1], *Spyros Matsoukas* [2], *Pavel Matějka* [1], *Najim Dehak* [3], *Jeff Ma* [2], *S. Cumani* [1], *O. Glembek* [1], *H. Hermansky* [4], *S.H. Mallidi* [4], *N. Mesgarani* [5], *R. Schwartz* [2], *M. Soufifar* [1], *Z.H. Tan* [6], *S. Thomas* [4], *B. Zhang* [2], *X. Zhou* [5]

[1]Brno University of Technology, Speech@FIT, Brno, Czech Republic
[2]Raytheon BBN Technologies, Cambridge MA, USA
[3]Massachusetts Institute of Technology, Cambridge MA, USA
[4]The Johns Hopkins University, Baltimore MD, USA
[5]University of Maryland, College Park MD, USA
[6]Aalborg University, Aalborg, Denmark

{iplchot,matejkap}@fit.vutbr.cz, {smatsouk,jma}@bbn.com, najim@csail.mit.edu

ABSTRACT

This paper describes the speaker identification (SID) system developed by the Patrol team for the first phase of the DARPA RATS (Robust Automatic Transcription of Speech) program, which seeks to advance state of the art detection capabilities on audio from highly degraded communication channels. We present results using multiple SID systems differing mainly in the algorithm used for voice activity detection (VAD) and feature extraction. We show that (a) unsupervised VAD performs as well supervised methods in terms of downstream SID performance, (b) noise-robust feature extraction methods such as CFCCs out-perform MFCC front-ends on noisy audio, and (c) fusion of multiple systems provides 24% relative improvement in EER compared to the single best system when using a novel SVM-based fusion algorithm that uses side information such as gender, language, and channel id.

Index Terms— speaker identification, noisy speech processing.

1. INTRODUCTION

The goal of the RATS program is to create technology capable of accurately determining speech activity regions, detecting key words, and identifying language and speakers, in highly degraded, weak and/or noisy communication channels. The data sets used in RATS are obtained by retransmitting pre-existing or newly collected telephone conversations in multiple languages over various types of channels, and aim to capture/simulate the acoustic environment present in current radio-based two-way communications systems used by the law enforcement, emergency, air traffic control, etc.

By its nature, these radio means of communication are sensitive to many factors which can degrade or change the quality of the transmission. The most important are background radio interference, atmospheric conditions, used bandwidth and background additive noise. All of these factors greatly increase the unwanted channel variability present in the audio.

When working with this type of data we are facing many new challenges and need to develop new techniques to overcome the problems caused by the channel distortions. We need to revisit every step of the technology used in current state-of-the art systems, which are mainly designed for much cleaner telephone conversations or interviews recorded with high-quality microphones in relatively low-noise environment.

In the speaker recognition system described in this paper, we had to begin by developing noise-robust models for voice activity detection based on both supervised and unsupervised methods. In addition, we experimented with various types of acoustic features in order to see their effect on the behavior of the system under noise. Finally, we explored alternative techniques for combining the scores of multiple subsystems - based on state-of-the-art speaker identification technology built on top of the i-vectors[1, 2] and probabilistic linear discriminant analysis (PLDA)[3] - using side information automatically extracted from the audio, such as gender and channel id.

The paper is organized as follows: Section 2 describes the training, development, and evaluation data sets. Section 3 explains the system components. Section 4 covers the fusion and calibration. Section 5 summarizes the results of individual systems as well as fusion, and section 6 ends the paper with discussion and conclusions.

2. DATA

The Linguistic Data Consortium (LDC) provided the training and test data for the RATS participants. The audio recordings were selected from existing and new data sources as follows: **NIST SRE 2004** (Eng., Ara., Chin., Rus., Span.); **RATS-LDC** (Lev. Arabic, Farsi); **RATS-Appen** (Lev. Ara., Farsi, Pash., Dari, Urdu); **Call-Friend Farsi**; **Fisher** (Lev. Ara. and Eng.); **NIST LRE** (various languages).

All recordings were retransmitted through 8 different noisy communication channels, labeled by the letters A through H [4]. A "push-to-talk" (PTT) transmission protocol was used in all channels except G. PTT states produce some regions where multiple non-transmission (NT) segments may occur. As a result, the amount of usable audio decreases after retransmission.

This work was partly supported by the DARPA RATS Program under Contract No. D10PC20015, by Technology Agency of the Czech Republic grant No. TA01011328, and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

It should be noted that among the data sources listed above, only the first three were annotated with speaker labels. Data from the other sources was used to train universal background models and i-vector extractors. We used the "dev" subset of the RATS-LDC and RATS-Appen corpora [1] to define speaker enrollment and testing samples. The rest of the RATS-LDC and RATS-Appen data, along with the NIST SRE 2004 set was used for speaker modeling.

There is also a separate blind "progress" test set, which is used to measure year-to-year progress on the RATS SID task. The progress set consists of speakers from the 5 target languages (Levantine Arabic, Farsi, Pashto, Dari, Urdu). Each speaker has 10 recording sessions, retransmitted over the 8 noisy channels as described above. For each speaker, 6 of the sessions are used for enrollment and 4 for testing, randomly sampled from the noisy channels. The progress set defines multiple testing conditions, depending on the amount of speech present in enrollment and testing samples. The following test-enroll conditions are evaluated (numbers indicate nominal amount of speech in seconds): 120-120, 30-30, 30-10, 30-3, 10-10, 10-3, 3-10, 3-3.

Only recordings from the 120s condition were released for training and development. We therefore had to construct our own development samples for the shorter durations from the 120s audio files, based on BUT's voice activity detection (VAD).

3. SYSTEM COMPONENTS

3.1. Voice activity detectors

NN-based VAD (VAD1): Voice activity detection is performed by Neural Network with input consisting of a block of Mel filter outputs with context of 300ms. The NN has 18 outputs: 9 for speech and 9 for non-speech, each corresponding to one of the channels (source plus 8 re-transmitted). HMM with Viterbi decoding is used to smooth out and merge the outputs to speech and non-speech regions. This NN is trained on RATS data defined for the speech activity detection (SAD) task [5].

GMM-based VAD (VAD2): This system is a variant of the GMM-based VAD described in [5]. The audio bandwidth is set to 125-3750Hz. Normalized energy and 14 perceptual linear predictive (PLP) coefficients are first extracted for every 25ms with a shift of 10ms. RASTA-based [6] normalization is applied to the PLP coefficients. The 15-dimensional feature vector at each frame is augmented with the corresponding features from the preceding 7 and following 7 context frames, and then projected down to 45 dimensions using heteroscedastic linear discriminant analysis (HLDA). Two 2048-component GMMs (speech/non-speech) were trained on the resulting feature space so as to maximize the mutual information between the training observations and their respective speech/non-speech labels.

Unsupervised VAD (VAD3) and denoising: Speech signals in RATS are corrupted by both relatively stationary noise as well as burst-like noise. Due to the very different characteristics of the two types of noise, they should be dealt with separately. Therefore, we investigated a two-pass segment-based method for VAD and denoising. In the first pass, the speech signal is first filtered by a first-order high-pass filter with a cutoff frequency of 60Hz. Then high energy segments are detected by using the a-posteriori signal-to-noise-ratio (SNR) weighted energy difference measure [7]. If the a-posteriori SNR weighted energy distance of two consecutive frames is larger than a predefined threshold, a high-energy frame is detected. Within a high energy segment, if no pitch is found, the segment is considered as noise. In this work, pitch detection is realized by using Praat software [8]. In the second pass, the speech signal is denoised by minimal statistics noise estimation (MSNE) based method to remove relatively stationary noise [9]. We used a modified version of MSNE adjusted for the RATS data. The final speech signal is denoised by setting the high-energy noise segment to zero. VAD is conducted on the denoised data. Pitch information is also used in this step on the assumption that all speech segments should contain pitch. The a-posteriori SNR weighted energy difference measure is applied now to the voiced speech segments to make voice activity detection.

[1] LDC catalog ids: LDC2012E49, LDC2012E63, LDC2012E69.

3.2. Acoustic and Prosodic Front-ends

In order to improve robustness to noise, we investigated various types of acoustic front-ends, described below.

MFCC: This front-end operates on standard Mel-frequency Cepstrum Coefficients (MFCC), extracted using a 25ms Hamming window. We extract 19 MFCCs together with log-energy every 10ms. We augment the MFCC with delta and double delta coefficients calculated using a 5 frame window, which results in 60-dimensional feature vectors. These are subjected to feature warping [10] using a 3s sliding window before removing the silence.

PLP: We bandlimit the audio to the 125-3750Hz range and extract 14 PLP coefficients plus normalized energy using a 25ms Hamming window with a 10ms frame shift. We augment the PLPs with their first and second derivatives, yielding 45-dimensional feature vectors, which are then subjected to feature warping using a 3s sliding window over the detected speech regions.

CFCC: We use auditory motivated features which simulate the signal processing functions in cochlea [11]. We use 24 Gammatone filters with frequency band 300-3400Hz. The resulting output is filtered through a low-pass filter with cutoff frequency 20Hz. Instead of using a fixed length window, we use a variable length window for different frequency bands. The higher the frequency, the shorter the window. This avoids the high frequency information being smoothed out by a long window duration. The window length is proportional to center frequency of the Gammaton filter. We apply the hamming window, take the logarithm, and apply discrete cosine transform (DCT) on the resulting window with 20 basis. We add deltas and double deltas, resulting in a 60-dimensional feature vector. Afterwards we remove silence frames according to VAD and we apply feature warping with a window of 3s.

FDLP: Auto-regressive (AR) modeling emphasizes the peaks of the spectrum, which are more salient. Frequency domain linear prediction (FDLP) extends this idea to model the time domain Hilbert envelope of the signal [12, 13]. The emphasis is on temporal peaks, which are more robust to noise. In this approach, we first apply DCT on 10s speech segments [14]. The full-band DCT is windowed into 96 linear sub-bands in the frequency range of 125-3800Hz. Linear prediction is performed on each sub-band to obtain parametric sub-band envelopes, which are then stacked to form a two-dimensional time-frequency representation, similar to spectrogram. This representation is decimated to 100Hz sampling rate, providing an estimate of the power spectrum of the signal in the short-term frame level. These linearly spaced power spectral estimates are then warped to mel axis by critical band integration [15], using a 3s sliding window, and they are finally converted to 60-dimensional features containing 20 cepstral coefficients along with their first and second derivatives.

PLP2: The output power spectral estimates from the critical band integration stage of FDLP, are inverse Fourier transformed to obtain an autocorrelation sequence [16]. This sequence is used for

time-domain linear prediction (TDLP), using a 19th-order model. The TDLP provides an all-pole approximation of the short-term spectrum. The output TDLP parameters are converted to 20 cepstral coefficients using cepstral recursion. Deltas and double-deltas are appended to generate a 60-dimensional feature vector at each time frame. Before removing the silence, feature vectors are warped using a 3s sliding window [15].

Cortical Features: The cortical representation of speech is derived from a two-stage computational auditory model [17], which is based on neurophysiological investigations of the human auditory system. The output of the auditory model is a multidimensional array of temporal and spectral modulations along time, frequency, rate, and scale. It is averaged over a 250 ms sliding window. We first reduce the high dimensionality of cortical features using a traditional principal component analysis (PCA) to 19 features. Then we compute and concatenate the delta and double-delta features to produce a 57-dimensional vector for each frame. Feature warping is applied next, using a 3s sliding window over the speech segments detected by VAD. The resulting features have been shown to have some robustness to additive noise and reverberation in the case where the speaker models are trained from clean data [18].

Prosodic Front-end: The prosodic system is trained over F0 and energy contours as the preliminary features. The F0 and energy of the signal are extracted using 10ms frames using the Snack toolkit [19]. The same VAD as in VAD1 is used. The F0 and energy contours are then estimated using a fixed length window of 200ms with 50ms shift. The contours are estimated using discrete cosine transform and the first 6 coefficients are used as the representative of the corresponding contours in each window. A 13-dimensional feature vector (6 F0 coefficients, 6 energy coefficients and number of the voiced frames) is then used to train a gender independent 2048-component UBM using diagonal covariance matrix. A 300-dimensional total variability subspace is then trained for extraction of the ivectors [20].

3.3. Modeling

Features resulting from the various combinations of voice activity detectors and acoustic/prosodic front-ends were used to train i-vector based SID systems. Three types of i-vector sub-systems were used, developed at BUT, MIT, and BBN. In the rest of the paper we will be referring to these sub-systems as ivec1, ivec2, and ivec3, respectively.

Common framework for training and scoring: A universal background model (UBM) is first trained, and first and second order statistics are extracted for every signal to be processed. The statistics from the training data are then used to train i-vector extractor which is then applied on all enrollment and test sessions to transform them into fixed-length low dimensional i-vectors. All subsystems included in our submission use the i-vector/PLDA framework for modeling. The i-vectors are transformed using linear discriminant analysis (LDA) and normalized to unit length. Log-likelihood ratios for each trial are estimated using probabilistic linear discriminant analysis (PLDA) [3]. The LDA transform and PLDA parameters are learned from i-vectors extracted from the training data.

Universal background model: Each sub-system used its own gender-independent universal background model (UBM), represented as a diagonal covariance Gaussian mixture model (GMM). Variance flooring was used in each iteration of EM algorithm during the UBM training. The UBMs for ivec1 and ivec2 had 2048 mixture components, while ivec3 had 1024.

i-vector extraction: The UBMs were used to generate zero and first order statistics for training the i-vector extractors [1, 2]. Subsystems ivec1 and ivec2 output 600-dimensional i-vectors, while sub-system ivec3 outputs 500-dimensional i-vectors.

4. SYSTEM CALIBRATION AND COMBINATION

We used two different approaches to our fusion and calibration. First approach is a classical and well-tested fusion using logistic regression and only the scores of the subsystems as inputs. The second one uses a support vector machine (SVM) with linear kernel and the inputs are scores of the subsystems as well as other side-information which is known or can be automatically extracted at test time.

Because of the lack of an independent held-out calibration data set, we used jack-knifing and we divided our development database into two independent parts on which we trained the parameters. Parameters trained on first part were applied to the second part and vice versa. When we finished the system development, we used all of the development data for the fusion without jack-knifing.

4.1. Logistic Regression Fusion

We use the freely available Bosaris toolkit [21], which provides a logistic regression solution for the calibration and fusion. Both calibration and fusion are based on the mapping:

$$l_t = a + \sum_{i=1}^{N} b_i s_{it}$$

where l_t is the fused (if $N > 1$) and calibrated output log-likelihood-ratio for trial t; N is the number of subsystems to be fused (if $N = 1$, then the result is just calibration); s_{it} is the score of subsystem i for trial t. The parameters to be optimized are the scalar offset a and the scalar combination weights b_i. These are optimized using logistic regression, which minimizes the cross-entropy between the scores and the *supervised calibration database*.

4.2. SVM Fusion

We also investigated the use of an SVM for fusion. Besides the individual SID system scores, the SVM classifier can also take in other measured features of the input audio, such as channel id, gender id, etc. We experimented with different types of SVMs and found the best results when we used ROC area as the objective function to maximize. ROC area is a performance measure defined as the fraction of pairs of positive and negative examples that are ranked in correct order:

$$\text{ROC Area} = 1 - \frac{\text{num. swapped pairs}}{\text{num. pos.} \times \text{num. neg.}}$$

A swapped pair is one where the positive sample has a lower score than the negative sample. The input to the classifier is a pair of one positive sample and one negative sample. The output of the classifier is 1 if the positive sample's score is higher, and -1 otherwise. The target output is always 1. Joachims [22] shows that there is an efficient way to perform such optimization.

We experimented with different SID systems and features as input to the SVM. The following "side information" features were considered: gender id; language of trial (Pashto, Levantine, other); test channel id (A-H); and number of times the test channel was seen in enrollment (0, 1, 2+ times). Note that the language id was provided to the systems at enrollment/test time. All other features were automatically extracted from the audio.

5. RESULTS

During our development for the RATS Phase 1 evaluation, we built several systems, differing in the VAD algorithm, acoustic front-end (one of the front-ends was prosodic), and i-vector extraction. All systems were trained using the data described in Section 2, and were evaluated on our development set in terms of equal error rate (EER), as well as in terms of the two RATS Phase 1 SID performance metrics, which were (a) the miss rate (Miss) at the target false alarm rate of 4%; and (b) the false alarm rate (FA) at the target miss rate of 10%. Table 1 shows the performance of the individual systems that participated in the final fusion experiments, in terms of the above metrics. The scores were obtained by pooling trials across all 8 channels.

It can be seen that the best results are obtained using CFCC features, VAD1, and ivec1 extractor. In an earlier set of experiments, shown in the first two rows of Table 2, we found that CFCCs provided superior performance across all channels, compared to using MFCCs. The third row in Table 2 shows the result we obtain with CFCCs when data from channel B are excluded from the PLDA model training. The EER on channel B increases from 7.7% to 9.8%, while staying about the same on all other channels. This indicates that although CFCCs are more robust than MFCCs, the system performance is still very sensitive to new channels.

Comparing systems 3 and 7 in Table 1 shows that the unsupervised VAD is very competitive to the supervised NN-based VAD in terms of the downstream SID performance. Looking at systems 7 and 9, we see that using the denoised audio hurts SID performance. The denoising technique in system 9 is based on the unsupervised VAD (VAD3). However in order to build the speaker verification system, we used VAD1. This mismatch between the two VADs may explain the degradation from the denoising technique.

Table 3 shows a comparison between alternative fusions, differing in the configuration (which systems get combined), algorithm (logistic regression vs. SVM), and usage of side information. We found no benefit for using side information in LR-based fusion, while such information helps when using the SVM. The results of Table 3 show that the SVM is better than LR, especially when combining a large number of systems (configuration B). These results are confirmed on the RATS progress set, as shown in Table 4, across all test-enroll duration conditions.

6. DISCUSSION AND CONCLUSIONS

In this paper, we described the patrol team submission for DARPA-RATS Phase 1 speaker identification evaluation, using audio from highly degraded communication channels. Our submitted system was a fusion of four sub-systems, which differ primarily in terms of the features and voice activity detection. It is already known [23] that fusing multiple sub-systems, which are similar in speaker modeling and different only in terms of VAD and features helps to improve the performance. Also, in previous work on language identification (LID) for the RATS project [24], we had observed that detection accuracy was very sensitive to the VAD employed. For this reason we built three different VADs. The first two VADs were supervised, based on NN and GMM modeling. The last one is a two pass unsupervised VAD based on denoising technique. The results show that the unsupervised VAD achieved similar performance compared to the supervised ones. We also used different features comprised by a variant of cepstral, cortical and prosodic information. We observed significant improvement (24% relative) in EER from combining multiple systems using a novel SVM-based fusion algorithm that benefited from side information such as gender, language, and channel id.

In future RATS evaluations, we will need to address the problem of making our systems more robust to unseen channels. We started studying this problem in this paper where we show the results of a single system by excluding one target channel from the training data. The results show that our systems are very sensitive to the unseen channel, even when using state of the art noise-robust features. The intuition is that the channel characteristics are very different and we should address this issue by developing model adaptation or audio de-noising techniques. We also believe that the unsupervised VAD that we developed will be more robust to unseen channels compared to the supervised VAD methods.

	FEA	VAD	SubSys	FA	Miss	EER
1	CFCC	VAD1	ivec1	2.4	6.8	5.3
2	FDLP	VAD1	ivec1	3.4	8.9	6.1
3	MFCC	VAD1	ivec2	2.6	7.2	5.4
4	PLP2	VAD1	ivec1	3.2	8.4	5.9
5	PROSO	VAD1	ivec1	22.6	41.3	15.7
6	CORT	VAD2	ivec3	3.4	8.8	6.1
7	MFCC	VAD3	ivec2	2.8	7.7	5.6
8	PLP	VAD2	ivec3	4.4	10.8	6.7
9	MFCC	VAD1*	ivec2	3.2	8.4	5.9

Table 1. Subsystems using various VADs and feature front-ends. Results (%) are given on our DEV set 30-30 test-enroll condition. System 9 uses VAD1, but with denoised audio based on VAD3.

Feature	A	B	C	D	E	F	G	H
MFCC	8.3	8.2	8.4	9.0	9.4	7.4	5.4	12.1
CFCC	7.4	7.7	7.8	8.9	8.3	6.5	5.4	11.6
CFCC-noB	7.3	9.8	8.0	8.8	8.2	6.6	5.3	11.7

Table 2. MFCC and CFCC front ends across channels A-H. VAD1, ivec1 sub-system. Results (% EER) on DEV 30s condition.

Config	Fusion type	SideInfo	FA	Miss	EER
A	LR	No	1.4	5.0	4.5
A	SVM	No	1.2	4.5	4.3
A	SVM	Yes	1.1	4.3	4.2
B	LR	No	1.3	5.0	4.5
B	SVM	No	1.2	4.3	4.1
B	SVM	Yes	1.0	4.0	4.0

Table 3. System combinations. Configuration 'A' refers to the combination of systems (1,3,6,7) in Table 1. Configuration 'B' refers to the combination of systems 1 through 9. Results (%) are given on our DEV set 30-30 test-enroll condition.

Fusion	120-120	30-30	30-10	30-3
Primary	5.4	8.6	13.1	19.8
Contrastive	4.7	8.6	10.6	16.8
Fusion	**10-10**	**10-3**	**3-10**	**3-3**
Primary	20.2	26.3	35.8	48.3
Contrastive	17.0	21.5	31.9	40.7

Table 4. Results on progress set. 'Primary Fusion' and 'Contrastive Fusion' refer to fusion configurations "A-LR" and "B-SVM", as described in Table 3, rows 1 and 6. Results are miss rate (%) at the RATS Phase 1 target false alarm rate of 4%.

7. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, 2010.

[2] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "I-vector extraction simplified," in *submitted to ICASSP*, 2011.

[3] S. J. D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.

[4] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *ISCA Speaker Odyssey*, 2012.

[5] Tim Ng, Bing Zhang, Long Nguyen, Spyros Matsoukas, Karel Vesely, Pavel Matějka, Xinhui Zhu, and Nima Mesgarani, "Developing a speech activity detection system for the darpa rats program," in *Proc. of Interspeech 2012*, Sept. 2012.

[6] H. Hermansky, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[7] Z. Tan and B. Lindbeg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *Journal of Selected Topics in Signal Processing*, pp. 798–807, 2010.

[8] P. Boersma and D. Weenik, "Praat: doing phonetics by computer (v. 5.1.05)," http://www.praat.org.

[9] Rainer Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[10] S. Sridharan J. Pelecanos, "Feature warping for robust speaker verification," in *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 213–218.

[11] Qi Li and Yan Huang, "Robust speaker identification using an auditory-based feature," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4514–4517.

[12] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *JASA*, vol. 105, pp. 1912–1924, 1999.

[13] M. Athineos and D. Ellis, "Autoregressive modelling of temporal envelopes," *IEEE Trans. of Signal Processing*, vol. 55, pp. 5237–5245, 2007.

[14] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.

[15] S. Ganapathy, J. Pelecanos, and M.K. Omar, "Feature normalization for speaker verification in room reverberation," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 4836–4839.

[16] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-d autoregressive models for speaker recognition," in *ISCA Speaker Odyssey*, 2012.

[17] Taishih Chi, Powen Ru, and Shihab A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *JASA*, pp. 887–906, 2005.

[18] Sridhar Krishna Nemala, Dmitry N. Zotkin, Ramani Duraiswami, and Mounya Elhilali, "Biomimetic multi-resolution analysis for robust speaker recognition," *EURASIP J. Audio, Speech and Music Processing*, vol. 2012, pp. 22, 2012.

[19] "http1://www.speech.kth.se/snack," .

[20] Marcel Kockmann, *Subspace modeling of prosodic features for speaker verification*, Ph.D. thesis, Brno, CZ, 2012.

[21] "https://sites.google.com/site/bosaristoolkit," .

[22] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. of the International Conference on Machine Learning (ICML)*, 2005.

[23] N. Brummer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2072–2084, 2007.

[24] P. Matejka, O. Plchot, M. Soufifar, O. Glembek, L. D'Haro, K. Vesely, F. Grezl, J. Ma, S. Matsoukas, and N. Dehak, "Patrol team language identification system for DARPA RATS P1 evaluation," in *Proc. of Interspeech 2012*, Sept. 2012.