Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

SUPRA-SEGMENTAL FEATURE BASED SPEAKER TRAIT DETECTION

*Gang Liu, John H.L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA
{Gang.Liu, John.Hansen}@utdallas.edu

ABSTRACT

It is well known that speech utterances convey a rich diversity of information concerning the speaker in addition to related semantic content. Such information may contain speaker traits such as personality, likability, health/pathology, etc. To detect speaker traits in human computer interface is an important task toward formulating more efficient and natural computer engagement. This study proposes two groups of supra-segmental features for improving speaker trait detection performance. Compared with the 6125 dimension features based baseline system, the proposed supra-segmental system not only improves performance by 9.0%, but also is computationally attractive and proper for real life application since it derives a less than 63 dimension features, which are 99% less than the baseline system.

Index Terms—speaker trait, personality, likability, pathology, supra-segmental feature

1. INTRODUCTION

Speaker trait detection is the study of signals beyond the basic verbal message or speech. Automatic recognition of speaker traits could be useful in many daily applications, such as healthcare monitoring (psychological analysis), stress assessment, deception detection, education tutoring systems, etc.

Although some speaker traits, such as age and gender, stress [1-3], height [4], sleepiness [5] and others [6], have been explored, there are less seldom explored traits, such as personality, likability and pathology which warrant exploration. In study [7], a pilot exploration for personality detection was considered based on linguistic cues which mainly relied on text. Researchers recently showed that likability can be robustly detected from real-life telephone speech [8]. Pathologic speech detection, based on single phonemes, also acquired high accuracy [9]. This study further explores these aspects in a unified way according to the Sub-Challenges outlined in [10].

The first is personality detection, with Personality assessed along five dimensions (also known as the Big Five) as in [7]:

Openness to experience (intellectual, insightful);
Conscientiousness (self-disciplined, organized);
Extraversion (sociable, assertive, playful);
Agreeableness (friendly, cooperative);
Neuroticism (insecure, anxious).

In this study, each of five personality dimensions (OCEAN) is mapped into: *X* or *NX*, where *N* means “not”, $X \in \{O, C, E, A, N\}$.

In the Likability Sub-Challenge, the likability of a speaker’s voice needs to be assessed on a binary decision basis: *L* or *NL* (Likeable or Non-Likeable).

In the Pathology Sub-Challenge, the intelligibility of a speaker’s voice needs to be assessed on a binary decision basis: *I* or *NI* (Intelligible or Non-Intelligible).

This study, like other studies in data mining, involves feature representation and model classification. Although limitation from a given feature sometimes can be compensated to some degree by a discriminative backend modeling technology, a reasonable feature front-end always plays a vital role in the success of such applications. In the baseline system [10], a brute force feature set with 6125 dimensions is used for all three Sub-Challenges, where some discriminant features may thus be undermined by less informative ones.

In this study, we propose a speaker trait detection method based on supra-segmental features. The assumption behind this approach is that, when compared with a short windowed feature extraction approach (usually 20~30ms), the supra-segmental feature can capture a more global picture of speaker trait since it is reasonable to expect the speaker will exhibit the single trait in one short utterance (for example, less than 10 seconds).

This paper is organized as follows. Sec. 2 describes the three databases employed in this study. Sec. 3 outlines the baseline system and the adopted performance metric. The proposed supra-segmental feature extraction scheme is presented in Sec. 4. Backend systems are described next in Sec. 5. Experiments are reported in Sec.6 and research findings are summarized in Sec. 7.

2. DATABASE

This study uses the Speaker Personality Corpus (SPC) for Personality detection which consists of 640 French clips of audio files. The majority of the data are approximately 10 sec in duration. Non-native speaker judgment ratings are provided for the Big Five personality traits to ensure ratings are determined based purely on acoustic cues.

The Speaker Likability Database (SLD) is used for the speaker Likability Sub-Challenge. Participant raters were instructed to rate telephone recorded speech stimuli according to the likability of each stimulus, without taking into account sentence content or transmission quality. This data set is labeled either as ‘likable’ (L) class or ‘non-likable’ (NL) class.

NKI CCRT Speech Corpus is used for the speaker Pathology Sub-Challenge. Unlike the previous two corpora which were sampled at 8 kHz, this corpus is sampled at 16 kHz. The speech consists of recorded neutral Dutch text read before and after concomitant chemo-radiation treatment (CCRT) for inoperable tumors of the head and neck. The audio clips are assessed by native Dutch speakers, who are also speech pathologists. Every

* This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

sample is labeled as belonging to either the 'intelligible' (I) class or 'non-intelligible' (IL) class.

Further details regarding the corpus are given in [11-13][1].

3. BASELINE SYSTEM

The baseline system for this study employs front-end features extracted by the open source platform: *openSMILE* feature extractor toolkit [14]. Backend classifiers are implemented using the open source data mining platform: *WEKA* [15].

3.1. Feature extraction of baseline

The baseline features are extracted on a per-frame level. Each frame contains 64 low-level descriptors (LLD) including MFCC, Zero-Crossing Rate, etc. The final feature set is produced by computing frame-level static functionals (e.g., mean, deviation, max) across each of these LLD streams. These features are also called utterance features, since the functionals are computed across the entire utterance. The feature dimension is 6125 [10]. These mechanically produced high-dimensional features will not only potentially dilute the contribution of the more saline feature, but also render some high computation classifier impossible.

3.2. Backend classifier of baseline

A linear Support Vector Machine (SVM) trained with Sequential Minimal Optimization (SMO) is used as the backend classifier for the baseline system, which is robust against over-fitting in high dimensional feature spaces. This backend is abbreviated as SVM-SMO in the remainder of this study.

3.3. Performance measurement metrics

The unweighted average (UA) recall is used to measure performance. In the binary case ('*X*' and '*NX*'), it is defined as:

$$UA(X) = \frac{\text{Recall(X)} + \text{Recall}(N\,\text{X})}{2} \qquad (1)$$

Our study relies on unweighted average recall rather than weighted average (WA) recall ('conventional' accuracy) since it is also meaningful for highly unbalanced data.

4. SUPRA-SEGMENTAL FEATURE

Frame-based features may be ideal for content-sensitive applications, such as automatic speech recognition (ASR) [38]. In speaker trait detection where content plays a less informative role, supra-segmental features should be more discriminative. Two groups of supra-segmental features are investigated here.

4.1. Shifted Delta Cepstrum

The first group is shifted delta cepstrum (SDC) features. The inclusion of SDC in the context of speaker trait detection can extract longer temporal information. It is reasonable to expect that speaker have a single trait in a relative broad time span. Compared with traditional dynamic procedure (for example, delta and double delta in MFCC), SDC can extract acoustic information beyond the word boundary. The SDC is in fact a *k* block of delta cepstrum coefficients and illustrated in Figure 1 [17, 37-42]. Suppose the basic set of cepstrum coefficients,{$c_j(t)$, j=1,2,..,N-1}, is available at frame t, where j is the dimension index and N the number of cepstrum coefficients, then the SDC feature can be expressed as

$$\begin{aligned} s_{(iN+j)}(t) &= c_j(t+iP+d) - c_j(t+iP-d), \\ i &= 0,1,\ldots,k-1 \end{aligned} \qquad (2)$$

where d is time difference between frames for spectra computation, P is time shift between each block, and k is the total number of blocks. SDC coefficients can be concatenated with the basic static cepstrum coefficients. Thus, we can obtain a feature vector by concatenating $c_j(t)$ and $S_{(iN+j)}(t)$ (j=0,…,N-1; i= 0, …,k-1), which is the SDC version of features. The classical SDC configuration *N-d-P-k* in language identification 7-1-3-7 (the overall dimension is 56) is adopted in this study (Though optimum performance is possible with other settings, we will not divert away from the main goal). It is noted that these features based on the basic set of frame-based cepstrum coefficients. Therefore, we call them frame-based supra-segmental features. In this study, state-of-the-art Power-normalized cepstral coefficients (PNCC) are used as static features [18] to derive supra-segmental features.

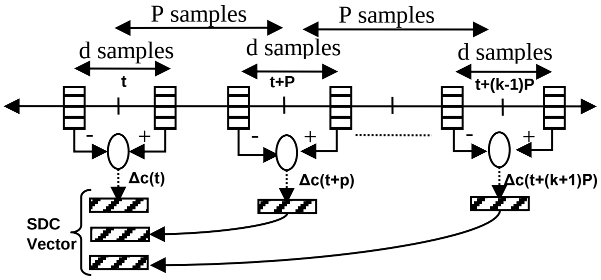

Figure 1. *Computation of the SDC feature vector at frame t for parameters N-d-P-k. The horizontal hatched box means the basic cepstrum coefficients, diagonal hatched box delta feature vector.*

4.2. Phoneme Statistics feature

The second group of supra-segmental features is based on phoneme statistics. High-level information, such as phoneme structure, usually carries some semantic cues. Some high-level characteristic, such as different speaker traits, inevitably has a direct impact on the production of speech and thereby the basic speech unit, phoneme. For example, people may elongate/ shorten some phoneme due to organs dysfunction. As a first step in this direction, a phoneme level emphasis will be investigated. However, one drawback for any phoneme approach is that a language dependent phoneme recognizer requires a significant amount of labeled phoneme transcription, which is time consuming and expensive. So, a language independent approach will be more practical. In this study where the speech involved French, German and Dutch, the independent Hungarian phoneme recognizer [19] is finally used to detect phoneme based features. Although there is language mismatch between phoneme recognizer and processed speech, the procedure followed here can be understood as sampling one language phoneme space with the codebook from another language phoneme space, and therefore, the phoneme recognizer can be used as speech unit detector/coder. This assumption will also be validated in the experiment stage. The phoneme statistics feature extraction is outlined as follows:

[1] Due to corpus license agreement, we do not have access to the test files' ground truth. All experiments are strictly following the training and development configuration as in [10].

Step 1: The phoneme recognizer first converts an acoustic utterance into a quadruple unit sequence. For example, the k^{th} phoneme quadruple unit in the i^{th} utterance can be coded as: (PHN_{ik}, BEG_{ik}, END_{ik} ,LLK_{ik}), where *PHN* is phoneme label, *BEG* segment beginning time, END segment ending time, and *LLK* log likelihood. *LLK* can be thought of as a measurement of the similarity between detected phoneme and phoneme behind the trained phoneme model. The larger the *LLK*, the more confident is the recognizer about its detection decision. This step therefore extracts the atomic phoneme feature.

Step 2: Calculate duration of the j^{th} phoneme and derive its mean and variance based on i^{th} utterance. This constitutes the duration feature stream: ($\mathrm{DUR}_{\mathrm{j\_mean}}$, $\mathrm{DUR}_{\mathrm{j\_var}}$). Derive the mean and variance of j^{th} phoneme LLK within each utterance. This constitutes the probability feature stream: ($\mathrm{LLK}_{j\_\mathrm{mean}}$, $\mathrm{LLK}_{j\_\mathrm{var}}$). *j* is in the range of [1, J], where J is the total phoneme number in the phone recognizer's dictionary.

Due to randomness from either speaker and/or speech contents, different phoneme statistics induce different impact to final performance. Therefore, we also need to find an optimal feature subset based on the basic unit from step 2. We propose the following three type feature subset:

DUR_mv_LLK_mv: a vector of concatenation of quadruple phoneme statistics unit ($\mathrm{DUR}_{\mathrm{j\_mean}}$, $\mathrm{DUR}_{\mathrm{j\_var}}$, $\mathrm{LLK}_{\mathrm{j\_mean}}$, $\mathrm{LLK}_{\mathrm{j\_var}}$). The vector dimension is 4J.

DUR_m: a vector of concatenation of unitary phoneme statistics unit ($\mathrm{DUR}_{\mathrm{j\_mean}}$) . The vector dimension is J.

DUR_m_LLK_avg: a vector of concatenation of DUR_m and *LLK_avg*, the average of {$\mathrm{LLK}_{\mathrm{j\_mean}}$}, *j* is in the range of [1,J]. The vector dimension is J+1.

After completing the above steps, each utterance is converted into a dimension-fixed feature vector which can be processed by using a frame independent classifier (such as an SVM-SMO).

5. BACKEND

Two groups of supra-segmental features are proposed in Sec 4. Due to their differences, two backends are investigated. Fusion is also explored to further improve performance. The entire proposed system is illustrated in Figure 2.

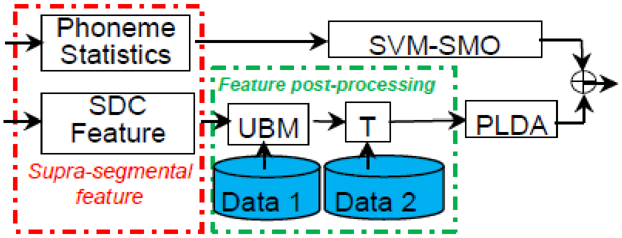

Figure 2. *Flowchart of the proposed system: Two groups of supra-segmental features + post-processing and backend fusion. In this study the data for UBM, Total variability matrix (T), and PLDA model development are the same as training data.*

5.1. i-Vector and PLDA classifier for SDC features

The first group is SDC-based supra-segmental features, which can also be called sub-utterance features and are still frame length-varied features. Therefore, we propose to adopt i-Vector and probabilistic linear discriminant analysis (PLDA) framework to fully explore the first group's discriminating capability. i-Vector and PLDA is the state-of-the-art framework for many speech-based identification tasks, such as identification of speaker[34, 35,44-47] and age [36] . i-Vector model is represented by

$$M = m + T\omega \tag{3}$$

where *T* is the total variability space matrix and ω is i-Vector, *m* is the UBM mean supervector, and *M* is the super-vector derived from supra-segmental features [20]. For each utterance, one i-Vector feature can be derived.

The i-Vector derivation procedure in this study is *i*) extracting supra-segmental acoustic features from each utterance, *ii*) grouping all the training data to train a universal background model (UBM), *iii*) computing Baum-Welch statistic for each utterance based on first two steps, *iv*) training total variability matrix *T* with all training data and extract i-Vector for both training data and test data. All these steps (after raw feature extraction and before classification) are noted as raw feature post-processing and illustrated in the green block of Figure 2. A 256-mixture UBM is trained for each task. A 50-dimension i-Vector is extracted for each audio file.

Note the matrix *T* contains both discriminative speaker trait information and non-speaker trait distortion information, Therefore, after extracting the i-Vector, PLDA is employed as the backend since it can effectively remove distortion [21]. Let the instance *j* of speaker trait *i* be ω_{ij} and let it be modeled as:

$$\omega_{ij} = Vy_i + Ux_{ij} + z_{ij} \tag{4}$$

where *V* and *U* are rectangular matrices and represent eigenvoice and eigenchannel subspace respectively. y_i and x_{ij} are the speaker trait factor and non-speaker trait factor respectively. The model parameters are learned from training data as each category has multiple utterances. Since the same speaker traits can be shared among different speakers and this study focuses on speaker independent trait detection, we expect better performance by removing the non-speaker trait distortion. During classification, the detection score is calculated like follows:

$$score(w_i \mid M_j) = \log \frac{p(w_i \mid M_j)}{p(w_i \mid \neg M_j)} \tag{5}$$

where M_j is the averaged i-Vector for the j^{th} speaker trait [22]. It should be noted that the i-Vector PLDA based framework gives better results than the raw acoustic feature based GMM [23,43]. Therefore, only results for the former are reported here.

5.2. SVM-SMO for phoneme statistics feature

The second group, phoneme statistics features, is utterance-wise features and has the same dimension number, though some dimensions may be missing due to varied speech contents. The SVM-SMO classifier in the baseline system is employed.

5.3. Fusion

Note that the i-Vector system can convert varied-length features into low dimensional fixed-length features and the SVM system can work with the fixed-length feature. To better explore discriminating capabilities of different front-ends and backends, linear fusion is deployed by using fusion toolkit Focal [24] (train data is used for fusion parameter learning).

6. EXPERIMENT RESULTS AND DISCUSSION

Based on the i-Vector and PLDA framework, the PNCC-SDC supra-segmental feature's performance on all of the three speaker trait challenges (personality, likability, and intelligibility), are illustrated in Figure 2. Except on personality A (Agreeableness) detection where the proposed SDC supra-segmental features are inferior to the baseline, they perform better in all the remaining six scenarios with significant gains. It should be noted that the dimension of i-Vector in this study is 50, which is far less than the baseline features' dimension, which is 6125. Although the i-Vector training stage is notoriously computationally demanding, it can be done off-line, which is beneficial to the on-line application.

The second group of supra-segmental feature, phoneme statistics features, should be more discriminative for phoneme-based tasks, such as detection of intelligibility versus that of personality or likability, in which the impact of phoneme variation is less prominent. Therefore, only the result for intelligibility detection is explored in this study. First of all, we want to find the optimal phoneme statistic features set. Performance of various phoneme features is summarized on Table 1. All proposed phoneme supra-segmental features can improve system performance, but the variance is less informative due to non-trait randomness and therefore is dropped in further exploration. Integration of averaged similarity indicator, LLK, can significantly boost performance by measuring how standard is the subjects' pronunciation, which in theory can help the intelligibility detection. Secondly, we want to prove the language-independent assumption behind our approach. Three phoneme recognizers, trained with Czech, Russian, and Hungarian languages respectively are used for the intelligibility detection. Dictionary size (or the phoneme count) of the three language are 45, 52, and 61. From Table 2, we can observe that, although the phoneme dictionary size varied significantly from one phoneme recognizer to another (maximal phoneme dictionary size varies by 35.6% relatively), performance varied only 1.4%. In addition, we should note that each phoneme recognizer has a different phonetic Alphabet. Therefore, based on the relative stable performance from Table 2, we can tentatively suggest that the phoneme statistic feature can robustly capture the intelligibility/non-intelligibility characteristics with the presence of language mismatch, therefore offering better scalability to be generalized to the unseen language in the field application. Another observation from Table 2 is that higher dictionary size can aid system performance since it results in higher resolution in the phoneme space. So, only the Hungarian phoneme recognizer is adopted thereafter.

To fully leverage the potential of the two supra-segmental features, fusion is applied and results are summarized in Table 3. Although phoneme statistics are a bit inferior to the SDC features, the 17.8% relative improvement against baseline proves complementarity of the two kinds of features.

Finally, we compared our proposed system with the baseline system and summarized results in Table 4. Across all the three speaker trait detections, the proposed system can consistently provide significant improvements and, compared with best results on the same experiment configuration, our system (noted as CRSS in Table 4) performs better on the Likability trait detection. Though admittedly the proposed system is inferior to the other two best published system, our proposed system can address the three speaker trait detection in a unified way.

7. CONCLUSIONS

This paper has described our efforts to detect speaker traits based on supra-segment acoustic features. The proposed SDC-iVector system can consistently improve performance across all the three speaker trait detection. Another group of novel phoneme statistics features also demonstrate their superiority on intelligibility detection and can dramatically improve system performance when fused with the SDC-based supra-segmental feature sub-system. Compared with the baseline system, the proposed system not only relatively improves performance by 9.0%, but also is computationally attractive and proper for real life application. It derives less than 63 dimension features, which are 99% less than the baseline system.

This study is based on the Speaker Trait Challenge 2012 corpora [10], which have already promoted some ongoing research. However, most of efforts focus on backend classifiers [25~28]. There are only a few researches involving feature, for example pitch and intonation [29], prosody [30], voice quality hierarchical feature [31], and this study targets expanding research on the trait dependent feature and also with real life application restrictions in mind, such as low-computation and scalability. It is also important to note that the proposed SDC-subsystem can consistently perform better than the baseline system, which is rarely the case in all the three best published trait challenge systems [32] [29] [33] since each of them either address only one sub-challenge or cannot perform better than the baseline system in all the three sub-challenges (i.e., Personality, Likability, and Pathology, respectively). This may suggest those systems are over-tuned for specific data.

8. ACKNOWLEDGEMENTS

The authors would like to thank Felix Weninger and Björn Schuller for their valued discussions and support.

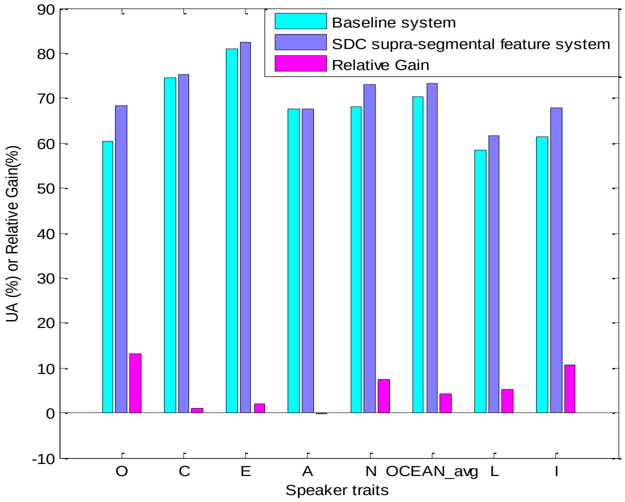

Figure 3. *Comparison between baseline system and proposed SDC supra-segmental system across seven trait detection tasks(OCEAN_avg is the average of five personality traits:OCEAN, L is Likability and I is intelligibility).*

Table 1. *Phoneme-based feature optimization on intelligibility detection. The performance measurement is UA(%). (m: mean; v: variance; The dimension of each feature type is put in parenthesis; Note, the phoneme recognizer has 61 phoneme units in the dictionary)*

Feature Scheme (Feature Dimension)	SVM
Baseline (6125)	61.4
DUR_mv_LLK_mv (61X4)	62.3
DUR_m (61)	63.8
DUR_m_LLK_avg (61+1)	**64.9**

Table 2. *Performance comparison of 3 phoneme recognizers on intelligibility detection.*

Phoneme Recognizer	Czech	Russian	Hungarian
Phoneme dictionary size	45	52	61
UA(%)	64.0	64.5	**64.9**

Table 3. *Fusion of phoneme statistics feature sub-system and SDC sub-system on intelligibility detection.*

Feature Scheme	UA(%)
Baseline	61.4
Phoneme Stats Feature	66.4
SDC feature	67.9
SDC system + Phoneme Stats system	**72.3**

Table 4. *Personality, Likability, and Pathology Sub-Challenge results on development dataset by baseline system and CRSS proposed system. The performance measurement is UA(%). Relative gain is computed as: (CRSS-Baseline)/Baseline X 100%.*

Task	Baseline	CRSS	Gain(%)	Best
OCEAN_avg.	70.3	73.3	+4.3	76.9[32]
(N)L	58.5	61.6	+5.3	61.1[29]
(N)I	61.4	72.3	+17.8	79.9[33]
Average	63.4	69.1	+9.0	/

9. REFERENCES

[1] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, “Nonlinear Feature-Based Classification of Speech Under Stress,” IEEE Trans. Speech & Audio Process., vol.9, no.3, pp.201-216, 2001.

[2] D. A. Cairns and J.H.L. Hansen, “Nonlinear analysis and classification of speech under stressed conditions,” Journal of the Acoustical Society of America, vol. 96, no. 6, pp. 3392–3400,1994.

[3] B. D. Womack and J.H.L. Hansen, “Classification of speech under stress using target driven features,” Speech Communication, vol. 20, no. 1-2, pp. 131–150, 1996.

[4] B. L. Pellom and J.H.L. Hansen, “Voice analysis in adverse conditions: the centennial Olympic park bombing 911 call,” in 40th Midwest Symp. on Circuits and Sys., 1997, pp. 873–876.

[5] T. Rahman, S. Mariooryad, S. Keshavamurthy, G. Liu , John H.L. Hansen, C Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features",INTERSPEECH-2011,pp.3285-3288.

[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. M üller, S. S. Narayanan, “The INTERSPEECH2010 Paralinguistic Challenge,” INTERSPEECH2010, Makuhari, Japan. pp.2822-2825.

[7] F. Mairesse et al. "Using linguistic cues for the automatic recognition of personality in conversation and text," Journal of Artificial Intelligence Research (JAIR), 30:457–500, 2007.

[8] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, “‘Would You Buy A Car From Me?’—On the Likability of Telephone Voices,” in Proc. of Interspeech. ISCA, 2011, pp. 1557–1560.

[9] A. A. Dibazar, S. Narayanan, and T. W. Berger, “Feature analysis for automatic detection of pathological speech,” in Proc. IEEE Joint EMBS/BMES Conf., Houston, TX, Oct. 2002, pp. 182–183.

[10] B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," INTERSPEECH 2012, ISCA, Portland, OR, USA.

[11] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, “A database of age and gender annotated telephone speech,” in Proc. International Conference on Language Resources and Evaluation (LREC). ELRA, 2010, pp.1562–1565.

[12] L. Molen, M. A. Rossum, A. H. Ackerstaff, L. E Smeele, C. R. N. Rasch and F. J. M. Hilgers, “Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients’ views,” BMC Ear Nose Throat Disorders, vol. 9, no. 10, 2009.

[13] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” Speech Communication, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[14] F. Eyben, M. W¨ollmer, and B. Schuller, “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in Proc. ACM Multimedia. Florence, Italy: ACM, 2010, pp. 1459–1462.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations,vol. 11, 2009.

[16] T. K. Ho, “The Random Subspace Method for Constructing Decision Forests,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 832–844, 1998.

[17] B. Bielefeld, "Language identification using shifted delta cepstrum," In 14th Annual Speech Research Symposium, 1994.

[18] C. Kim, R. M. Stern. “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” in Proc. ICASSP, Kyoto, Japan, pp. 4101-4104.

[19] P. Schwarz, "Phoneme Recognition based on Long Temporal Context, PhD Thesis", Brno University of Technology, 2009.

[20] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, “Front End Factor Analysis for Speaker Verification,” IEEE Transactions on Audio, Speech and Language Processing,pp.788-798. 2011.

[21] S. J. D. Prince, J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in Proc. ICCV, 2007, pp. 1–8.

[22] G. Liu, T. Hasan, H. Boril, J.H.L. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment", in Proc. ICASSP, Vancouver, Canada. 2013. pp. 7755-7759

[23] G. Liu, Y. Lei, J.H.L. Hansen, "A Novel Feature Extraction Strategy for Multi-stream Robust Emotion Identification", INTERSPEECH2010. Makuhari Messe, Japan. pp.482-485

[24] N. Br ümmer, “Focal multi-class—tools for evaluation, calibration and fusion of, and decision-making with, multi-class statistical pattern recognition scores.” Online on: http://sites.google.com/site/nikobrummer/focalmulticlass.

[25] Y. Attabi, P. Dumouchel, “Anchor Models and WCCN Normalization for Speaker Trait Classification,” INTERSPEECH 2012, ISCA, Portland, OR, USA. pp. 522-525

[26] D. Lu, F. Sha, “Predicting Likability of Speakers with Gaussian Processes,” INTERSPEECH 2012, ISCA, Portland, OR, USA. pp.286-289

[27] N. Cummins, J. Epps, J. M. K. Kua, “A Comparison of Classification Paradigms for Speaker Likability Determination,” INTERSPEECH2012, ISCA, Portland, OR, USA. pp.282-285

[28] K. Audhkhasi, A. Metallinou, M. Li, S. Narayanan, “Speaker Personality Classification Using Systems Based on Acoustic-Lexical Cues and an Optimal Tree-Structured Bayesian Network,” INTERSPEECH-2012, ISCA, Portland, OR, USA. pp.262-265

[29] C. Montacié, M. Caraty, “Pitch and Intonation Contribution to Speakers’ Traits Classification,” INTERSPEECH2012, ISCA, Portland, OR, USA. pp.526-529

[30] M. H. Sanchez, A. Lawson, D. Vergyri, H. Bratt, “Multi-System Fusion of Extended Context Prosodic and Cepstral Features for Paralinguistic Speaker Trait Classification,” INTERSPEECH2012, ISCA, Portland, OR, USA. pp.514-517

[31] D. Huang, Y. Zhu, D. Wu, R. Yu, “Detecting Intelligibility by Linear Dimensionality Reduction and Normalized Voice Quality Hierarchical Features,” INTERSPEECH2012, Portland, OR, USA. pp.546-549

[32] A. V. Ivanov, X. Chen, “Modulation Spectrum Analysis for Speaker Personality Trait Recognition”, INTERSPEECH2012, Portland, OR, USA. pp.278-281

[33] J. Kim, N. Kumar, A. Tsiartas, M. Li and S. S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple high level descriptors", INTERSPEECH2012, Portland, OR, USA. pp.534-537

[34] J. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin, and J. H. L. Hansen, "Exploring Hilbert envelope based acoustic features in i-Vector speaker verification using HT-PLDA," in Proc. NIST Speaker Recognition Evaluation, Atlanta, GA, USA, Dec. 2011.

[35] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. L. Hansen, “CRSS systems for 2012 NIST speaker recognition evaluation”, in Proc. ICASSP, Vancouver, Canada, May 2013, pp. 6783-6787.

[36] M. Bahari, M.H.and McLaren, H.V. Hamme, and D.V. Leeuwen, “Age estimation from telephone speech using ivectors,” INTERSPEECH2012, Portland, OR, USA. pp.506-509

[37] Q. Zhang, G. Liu, and J. H. L. Hansen, “Robust Language Recognition Based on Hybrid Fusion,” in Proc. Odyssey 2014, The speaker and language recognition workshop, Joensuu, Finland, June 2014

[38] G. Liu, D. Dimitriadis and E. Bocchieri, "Robust speech enhancement techniques for ASR in non-stationary noise and dynamic environments", in Proc. INTERSPEECH, Lyon, France, 25-29 Aug.,2013. pp. 3017-3021

[39] G. Liu, Y. Lei, J.H.L. Hansen, "Dialect Identification: Impact of difference between Read versus spontaneous speech", EUSIPCO-2010. Aalborg, Denmark, 2010. pp.2003-2006

[40] G. Liu, Y. Lei, J.H.L. Hansen, "Robust feature front-end for speaker identification", in Proc. ICASSP, Kyoto, Japan, pp.4233-4236, 2012.

[41] G. Liu, C. Zhang, J.H.L. Hansen, "A Linguistic Data Acquisition Front-End for Language Recognition Evaluation", in Proc. Odyssey, Singapore, pp. 224-228, 25-28 June 2012.

[42] G. Liu, J.H.L. Hansen. "A systematic strategy for robust automatic dialect identification", EUSIPCO2011, Barcelona, Spain, 2011. pp.2138-2141

[43] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.F. Chen, J. Li, B. Firner, "Crowd++: Unsupervised Speaker Count with Smartphones," The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM UbiComp), Zurich, Switzerland, September 9-12, 2013. pp.43-52.

[44] V. Hautamaki, K. A. Lee, D. Leeuwen, R. Saeidi, A. Larcher, T. Kinnunen, T. Hasan, S. O. Sadjadi, G. Liu, H. Boril, J.H.L. Hansen and B. Fauve, "Automatic regularization of cross-entropy cost for speaker recognition fusion", in Proc. INTERSPEECH, Lyon, France, 25-29 Aug.,2013.

[45] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P. M. Bousquet, E. Khoury, P. L. Sordo Martinez, J. M. K. Kua, C. H. You, H. Sun, A. Larcher, P. Rajan, V. Hautamaki, C. Hanilci, B. Braithwaite, R. Gonzales-Hautamaki, S. O. Sadjadi, G. Liu, H. Boril, N. Shokouhi, D. Matrouf, L. El Shafey, P. Mowlaee, J. Epps, T. Thiruvaran, D. A. van Leeuwen, B. Ma, H. Li, J.H.L. Hansen, J. F. Bonastre, S. Marcel, J. Mason, E. Ambikairajah, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification", in Proc. INTERSPEECH, Lyon, France, 25-29 Aug.,2013.

[46] C. Yu, G. Liu, S. Hahm, and J.H.L. Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition," in Proc. ICASSP, Florence, Italy, May 2014

[47] G. Liu, J.W. Suh, J.H.L. Hansen, "A fast speaker verification with universal background support data selection", in Proc. ICASSP2012, Kyoto, Japan, pp.4793-4796, 2012.