PHONETIC AND ANTHROPOMETRIC CONDITIONING OF MSA-KST COGNITIVE IMPAIRMENT CHARACTERIZATION SYSTEM

Alexei V. Ivanov, Shahab Jalalvand, Roberto Gretter, Daniele Falavigna

Fondazione Bruno Kessler
via Sommarive, 18, Povo (Trento), Italy
alexei_v_ivanov@ieee.org, jalalvand@fbk.eu, gretter@fbk.eu, falavi@fbk.eu

ABSTRACT

We explore the impact of speech- and speaker-specific modeling onto the Modulation Spectrum Analysis – Kolmogorov-Smirnov feature Testing (MSA-KST) characterization method in the task of automated prediction of the cognitive impairment diagnosis, namely dysphasia and pervasive development disorder. Phoneme-synchronous capturing of speech dynamics is a reasonable choice for a segmental speech characterization system as it allows comparing speech dynamics in the similar phonetic contexts. Speaker-specific modeling aims at reducing the "within-the-class" variability of the characterized speech or speaker population by removing the effect of speaker properties that should have no relation to the characterization. Specifically the vocal tract length of a speaker has nothing to do with the diagnosis attribution and, thus, the feature set shall be normalized accordingly. The resulting system compares favorably to the baseline system of the Interspeech'2013 Computational Paralinguistics Challenge.

Index Terms— speech characterization, feature selection, modulation spectrum

1. INTRODUCTION

Detection and differentiation between distinct types of speech production impairments is important in medical domain for the large scale inexpensive and noninvasive pre-screening of children for cognitive development disorders. Self- administered measurement of the severity of speech production impairments is deemed to be very helpful in compensatory therapy and symptom management of the cognitive deficiency as it is the least embarrassing for the patient. Automated spoken agents would also benefit from the ability to adjust depending on cognitive abilities of the user. Speech is a complex and rapid cognitive process, thus, one can expect that various types of cognitive deficiencies would manifest themselves differently in speech. It is reasonable to conjecture that articulation, as the most rapid of speech production processes, shall be among the most informative data sources. Articulation is also convenient because its proper measurement does not require involvement of overly complex, noisy and error-prone models (i.e. lexicon, syntax, semantics, pragmatics, rhetoric).

In the present experiment we use the data that has been distributed for Interspeech'2013 Computational Paralinguistics Challenge [1]. Specifically the task is to discriminate automatically between autism, non otherwise specified pervasive developmental disorder, aphasia (partial loss of speech function, termed as dysphasia) and examples of normal cognitive function [3],[4]. Speech is sampled as short non-spontaneous read (or repeated) complete sentences. The dataset is split by the organizers into training (903 utterances), development (819 utterances) and test (820 utterances) sets. Test set labels are not available to prevent model tuning for the specific test set. Each Challenge participant has at most five attempts to verify the generated test set labels by submitting them to the organizers via a web interface. The dataset contains instances of speech recordings from about a hundred young native speakers of French (age distribution between 6 and 18 years old).

As a starting point for the described experiment we use our existing universal speech characterization system [2]. The core MSA-KST method has proven to be an efficient tool to capture speech dynamics in a statistical model. Essentially the method consists of

- Over-generation of possible features in spectral – modulation-spectral domain (see Fig. 1) and creating an exhaustive description of a non-stationary source dynamics.
- Pruning the resulting feature space with a statistically-motivated criterion. Experimentally it has been found that the non-parametric Kolmogorov-Smirnov statistical test, applied at the level of individual features works as reasonable computationally efficient heuristics.

The versatility of the MSA-KST speech characterization method is provided by the ability to find useful features for empirically defined tasks. No prior theories about characterization are required to develop a characterization system for the particular task. In the present work we propose two essential ways to improve the original system of [2]:

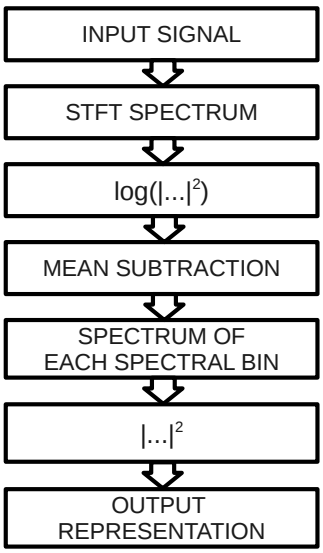

Fig. 1. Schematic structure of the MSA feature extraction.

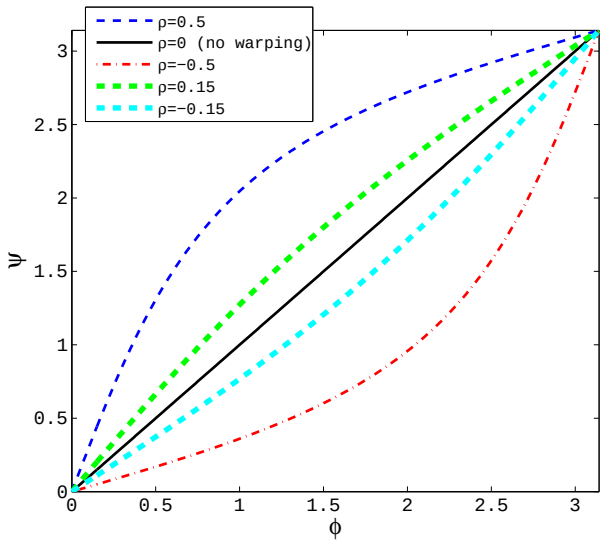

Fig. 2. Warped Discrete Fourier Transform (WDFT) scale alternation.

- Phoneme-type conditioned statistical models of speech dynamics (similar to what has been done in [5]). Our conjecture is that phoneme conditioning should aid robustness of the resulting model towards the phonetic content of the speech sample.

- Vocal Tract Length Normalization (VTLN) [6] of the data. The VTLN should be essential because the data in the Computational Paralinguistics Challenge and one of possible target application domains are specifically geared towards recognition of children's speech. VTLN is implemented in MSA domain via Warped Discrete Fourier Transform (WDFT).

The paper is organized as follows: second section discusses modification of the original MSA feature extractor, which enables it to compensate for the length of the vocal tract of a subject; third section explains our approach to phoneme-conditioned segmental modeling; fourth section outlines the experiments that have been performed in order to evaluate the presented modifications; conclusions are presented in the end of the paper.

2. VTLN IMPLEMENTATION IN MSA

There is a possibility to implement the VTLN procedure via a substitution of the conventional short time Fourier transform of the first stage of the modulation spectrum analysis (MSA) algorithm (see Fig. 1) with the WDFT [7, 8].

Frequency warping is a conformal map in the z-plane, which transforms the unit circle into itself. Particularly, a map, which is defined by the first order all-pass filter of the form:

$$\tilde{z}^{-1} = A(z) = \frac{\rho + z^{-1}}{\rho z^{-1} + 1}, \tag{1}$$

allows to warp frequency scale with the law :

$$tg\frac{\psi}{2} = \frac{(1-\rho)}{(1+\rho)} tg\frac{\varphi}{2}. \tag{2}$$

Here φ is a "linear" frequency scale (which constitutes to the unit circle in the z-plane), and ψ is a "warped" frequency scale in the $\tilde{z}$-plane by transformation (1). Frequencies ψ and φ are the normalized and, thus, do not depend upon the sampling rate.

With $0 < \rho < 1$ the transformation (1) stretches spectral bins in the low-frequency range and compress them in the high-frequency range. With $-1 < \rho < 0$ it does the other way around by stretching the high-frequency range and compressing the range at low-frequencies (see Fig. 2 for details).

Each speaker has his/her optimal value for the ρ parameter. With this speaker-dependent optimal ρ the transformation (1) of the frequency scale of MSA analysis will make the resulting feature set approximately invariant of the length of the speaker's vocal tract.

Learning a specific value of ρ that fits the speaker under consideration is done via matching the transformed speech against a general speech model. Below is a summary of the algorithm:

- Train a GMM general speech model from a large body of speech. Feature extraction method for this step is a classical set of first 13 cepstral coefficients drawn from the fully decimated output of a filter bank com-

plimented with estimates of their two first derivatives (39 features in total);

- Estimate the generation likelihood of the speech coming from a particular speaker according to the GMM model;
- Repeat the above procedure for a range of plausible values of ρ;
- Select the ρ value that gives the largest likelihood as an optimal value for the speaker and use it for feature extraction;
- The whole described procedure might be iterated further to make sure that the GMM in the first step has been trained on the speaker-invariant data.

The interval of $-0.2 < \rho < 0.2$ is sufficient to cover variation in adults and, as it has been experimentally found, is mostly sufficient for kids between 6 and 19 years of age. The range of estimated value in this experiment is $\rho \in [-0.01; 0.24]$.

The more speaker specific data is used in estimation of ρ the more precise and unambiguous the final result could be. Fig. 3 presents an exemplar outcome of such procedure. Here the speech evidence is sufficient to observe a clear maximum at $\rho = 0.06$.

As a sanity check for the parameter estimation procedure we have calculated the correlation between the parameter ρ and the reported age of a child. For the most successful configuration of the GMM model (one, containing 128 mixture components) such correlation is observed as high as 0.6740. This fact is encouraging as it is known from the statistics of child development that correlation between the age and height of a child is about 0.7 [9]. And length of the vocal tract has a very strong statistical relation with the height of a subject, e.g. a statistical analysis [10] of the relation of the vocal tract length and the height of subjects aged between 2 and 25 has revealed very high correlation (0.926).

Fig. 4 presents a comparative illustration of the MSA features computed with different values of ρ parameter. As it can be seen, with $\rho > 0$ the lower part of the spectrum gets stretched at the expense of compressing the higher part. If $\rho < 0$ the picture is inverse - the higher spectrum stretches while the lower gets compressed.

The training and development divisions are provided with a speaker label for each sentence. For the test, however, there is a need to automatically cluster utterances into groups, allegedly coming from the same speaker. In order to estimate the most plausible division of test utterances into speaker-specific clusters we generally follow the approach of iterative speaker normalization of [11].

In practice however, evaluation of the appropriate ρ shall not present a significant obstacle as an accurate measurement of the patient's height is readily available in most of the cases.

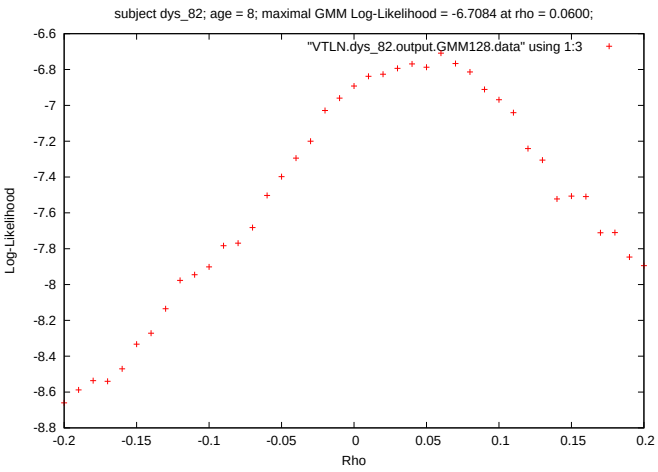

Fig. 3. Dependence of speech log-likelihood over the value of VTLN correction parameter ρ.

For telemedical applications the estimate of ρ can be drawn from the sufficient statistics of the patient's speech.

3. PHONEME-CONDITIONED MODELING

Phoneme-conditioned statistical models are implemented for broad phonetic classes, that follow five main manners of articulation:

- vowels {**a,e,E,i,o,O,u,@,9**}, which comprise 43.24% of total phonemes in transcription;
- plosives {**b,p,d,t,g,k**} occurring in 28.12% of cases;
- approximants, glides, trills and flaps {**w,y,l,j,R**} occurring in 10.64% of cases;
- nasals and nasalized vowels {**ã ,ẽ ,õ ,9̃ ,m,n**} occurring in 10.04% of cases;
- fricatives {**f,H,s,S,v,z,Z**} occurring in 7.95% of cases;
- the silence class, which sometimes contains vocalizations, that are far from any typical phoneme realization.

The underlying hypothesis for splitting a universal speech model into phone-specific parts is that pronunciation for different phonetic classes is affected in different ways and that the universal model is too crude of representation, which averages out some of the important characteristic distinctions. This conjecture is well grounded as instances of different phonetic classes have energy concentrated different regions in modulation-spectral domain. They also have different characteristic lengths, and are stretch-able to different extents. It is also reasonable to assume that consistency of the pronunciation alternations within phonetically-grouped classes is going to be higher.

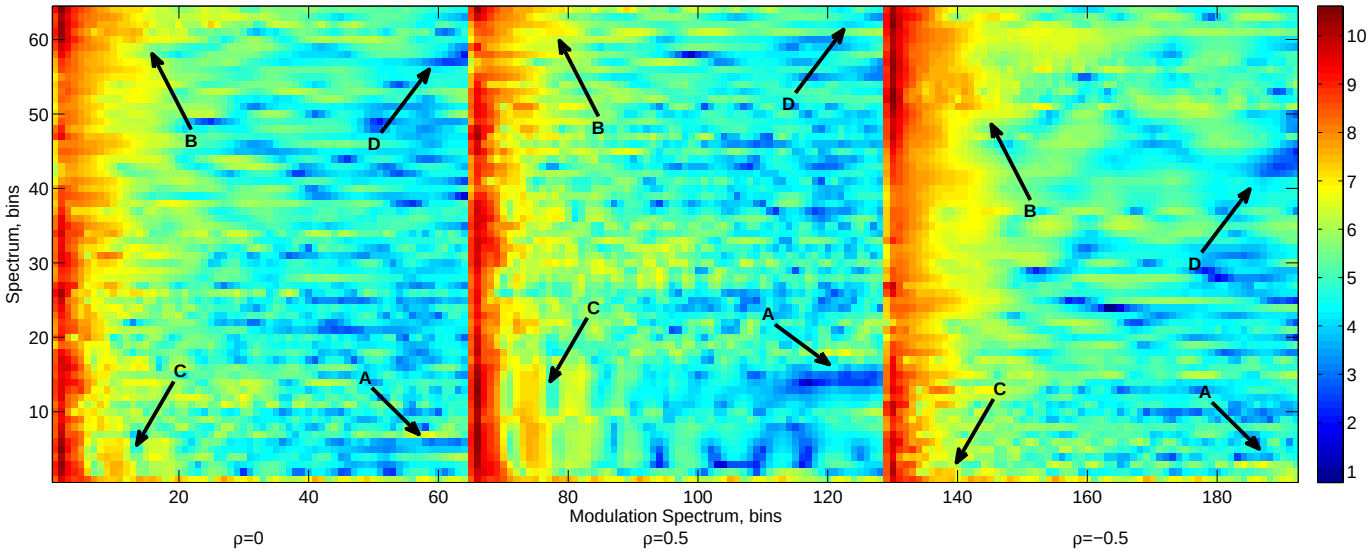

Fig. 4. Exemplar MSA feature set (log-variances of individual modulation spectrum bins through a speech sample).The plot is given for three values of ρ: ρ=0 (no distortion), ρ=0.5 (low spectral frequencies get stretched), ρ=-0.5 (high spectral frequencies get stretched). A, B, C, D - specific distinctive features on the spectral – modulation-spectral plane, which illustrate stretching and compression.

Phonetic transcription that is required to perform phonetic conditioning is obtained through force alignment of the available audio against the corresponding word string by the French ASR system that is developed at Fondazione Bruno Kessler. There are some cases where forced alignment fails completely because of the severe differences between the recorded audio and the intended sentence. In such cases we are backing off to the decision, provided by the system without phonetic conditioning.

For the test set we hypothesize that the data consists of the same set of 26 different sentences, that are found in the training and development part. This assumption follows from the database description, provided by the Interspeech'2013 Computational Paralinguistics Challenge organizers [1]. We do recognition of the provided audio with a phonetic network, that incorporates all 26 known sentences as recognition alternatives.

Each of the phonetically-conditioned classifiers is trained independently on the corresponding training data that is extracted from the total available material. At the recognition stage the likelihoods estimated by individual models are combined into the final decision of the system.

4. EXPERIMENTS

As in our previous system [2] we extract MSA features at 5 different rates with the analysis width being 8, 16, 32, 64 and 128 samples. The width of analysis window in modulation spectrum estimation always matches that from spectral estimation. Frame shift is always a half of the analysis frame. We convert spectral representation to log-power spectral domain and perform frame-mean subtraction before computing the modulation spectrum. The difference with the previous version however is in reduction of number of statistics computed over the sequence of raw features, this time only variance and kurtosis are computed. Thus, the whole set of MSA features plus the baseline set amounts roughly to fifty thousand different features before pruning.

KST pruning for the present experiment has been done as a selection of five thousand best features for the binary classification tasks separating each of the existing diagnoses from the rest. A combination of the selected sets is used as a KST prediction for the 4-way classification task.

In order to verify validity of the proposed feature extraction and selection strategies a recognition experiment has been performed. The recognizer is implemented as an adaptive meta-learning that aims at combining an ensemble of weak classifiers to form a strong classifier over one-level decision trees (Adaboosting) [12]. Specifically, an open source implementation "icsiboost" is used. Training is done for 600 iterations. The classifier is trained for a 4-way classification task to determine a diagnosis of the patient. The figures for the typicality classification are obtained by reassignment of the labels after a 4-way classification. This way we're not getting the top possible performance in typicality task as we have experimentally observed improvements by training a separate dedicated 2-way classifier.

Table 1 summarizes recognition results on the official Computational Paralinguistics Challenge evaluation sets (development and test). Along with the challenge organizers have provided their baseline system results [1], which are refered as "BL" in the table. Not all measurements are present for the official test set as only five attempts are allowed per Challenge participant. However, comparison of the figures achieved on the development and test portions indicate absence of overfitting by tuning towards the development set. Unweighted Average Recall (UAR) between $m = \overline{1..M}$ classification categories is chosen as the figure of merit according to the requirements of the Challenge [1]:

$$UAR = \frac{1}{M} \sum_{\forall m} \frac{N_{Corr_m}}{N_{Tot_m}}. \quad (3)$$

Here N_{Corr_m} is a number of correctly classified instances of the category m, N_{Tot_m} is a total number of instances of the category m.

As it can be seen from the table, KST pruning brings a significant improvement to the classifier performance, which is inline with our expectations as well as recognition experiment from [2]. VTLN does improve classification performance even better. Joint application of KST and VTLN rises performance further.

Employment of phonetically conditioned models also brings a significant improvement, although to somewhat lesser than expected extent. More detailed analysis of confusion matrix allows to see that majority of confusion occurs between autism (PDD) and pervasive development disorder non otherwise specified (NOS) cases. This behavior is very much expected as separation between the classical and non-specific autism types is the hardest task of all trained in the present experiment. The recall value for typically developing subjects is 93.55% and that for dysphasic subjects is 84.62%.

As anticipated, a combination of the VTLN and phonetic conditioning makes the system performance to go even higher. The VTLN and phonetic conditioning are aimed at improving different aspects of the feature extraction mechanism. Their effect on the information content of the features should have approximately additive effect.

Having more data from the subject allows to build much better classification systems. See, for example, results obtained for the configuration "+KST+VTLN+MV" and "+KST+PHO+MV". This is a situation when we applied label post-processing by majority vote within each speaker. Unfortunately we are not successful in deciphering speaker attribution for the test partition with automatic means. Our figures for the official test set UAR in "+KST+VTLN+PHO" were 61.78%, 61.55% and 61.62% depending on different tuning of speaker clustering procedure. The UAR averages for the "+KST+VTLN+PHO" case are presented in Table 1. However, in practical applications the true speaker identity is not hidden from the system.

Another interesting observation is an apparent presence of certain kind of label post-processing in the baseline figures for the Autism Challenge. Judging from the figures alone it appears that there is either a label post-processing or the classification task of the test set is much simpler that of the development set (UAR of $\sim 67\%$ vs $\sim 55\%$), however, our experience of submitting our results on the test set indicates that the test task is of similar complexity compared to the development case.

Table 1. *Performance of predictors of individual diagnosis (DIAG) - 4-way classification between autism a.k.a. Pervasive Development Disorder (PDD), Pervasive Development Disorder Non-Otherwise Specified (NOS), Dyspasia (DYS), Typical development (TYP) and detection of typicality of development (TYP-TY) - 2-way classification between typical (TYP) and atypical (ATP) development;"UAR" – unweighted average recall; "DEV" – development set;"TEST" – test set;"BL" – baseline;"MSA+BL" – a union of MSA & Baseline features; "+KST" – MSA+BL features, pruned with KST;"+VTLN" – MSA+BL features, obtained with vocal tract normalization;"+KST+VTLN" – MSA+BL features, obtained with vocal tract normalization & pruned with KST;"+KST+VTLN+MV" – MSA+BL features, obtained with vocal tract normalization & pruned with KST, a majority vote over individual utterances decides a diagnosis label for a given speaker;"+KST+PHO" – MSA+BL features, pruned with KST, recognizer utilizes phonetically-conditioned models;"+KST+VTLN+PHO" – MSA+BL features, obtained with vocal tract normalization, pruned with KST, recognizer utilizes phonetically-conditioned models;"+KST+PHO+MV" – MSA+BL features, pruned with KST, recognizer utilizes phonetically-conditioned models, a majority vote over individual utterances decides a diagnosis label for a given speaker.*

UAR	DEV		TEST	
System	DIAG	TYP-TY	DIAG	TYP-TY
BL	52.40	92.80	67.10	90.70
MSA+BL	**53.49**	85.57	–	–
+KST	**58.98**	89.28	60.24	88.66
+VTLN	**59.20**	88.20	–	–
+KST+VTLN	**62.67**	**94.60**	–	–
+KST+VTLN+MV	**75.37**	**100.00**	–	–
+KST+PHO	**62.29**	**92.97**	62.74	**92.41**
+KST+PHO+MV	**75.37**	**97.72**	–	–
+KST+VTLN+PHO	**68.26**	92.01	~ 61.65	$\sim$ **93.77**

5. CONCLUSIONS

We have obtained an experimental confirmation that proposed improvements to the MSA-KST speech and speaker characterization system are useful. Vocal tract length normalization

allows to make the feature set more homogeneous that enables better generalization from the training data. Phoneme-synchronous capturing of the speech dynamics is a reasonable choice for a segmental speech characterization system as it allows comparing segmental speech dynamics in similar phonetic contexts. Our automated diagnostic system for autism and dysphasia from speech is accurate given sufficient amount of material. Analysis of 26 spoken sentences is enough to make a nearly perfect separation of pathological speech and achieve $\sim 75\%$ UAR in differential diagnostics.

6. ACKNOWLEDGEMENTS

This work has been partially funded by the European project EUBRIDGE, under the contract FP7-287658.

7. REFERENCES

[1] Björn Schuller et all."The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism", Proc. of Interspeech'2013, Int. Conf., Aug. 25-29, 2013, Lyon, France

[2] Ivanov, A. and Chen, X., "Modulation Spectrum Analysis for Speaker Personality Trait Recognition", Proc. of Interspeech'2012, Int. Conf., Sept. 9-13, 2012, Portland, OR, USA

[3] "The ICD-10 for Mental and Behavioural Disorders Diagnostic Criteria for Research", World Health Organization, Geneva, Switzerland, 1992, 263 p.-

[4] "Diagnostic and Statistical Manual of Mental Disorders", 4th ed., American Psychiatric Association, 1994, 886 p.-

[5] M. Tang, "Large Vocabulary Continuous Speech Recognition Using Linguistic Features and Constraints", PhD Thesis, MIT, 2005, 123 p.-

[6] L. Lee and R. C. Rose, "Speaker normalization using efcient frequency warping procedures," Proc. ICASSP, vol. 8, no. 2, pp.353 - 356, May 1996.

[7] Makur, A. and Mitra, S.K., "Warped Discrete-Fourier Transform: Theory and Applications", IEEE Trans. On Circuits And Systems–I: Fundamental Theory And Applications, Vol. 48, No. 9, pp. 1086–1093, September 2001

[8] Ivanov, A.V. and Parfieniuk, M. and Petrovsky, A.A. "Frequency-Domain Auditory Suppression Modeling (FASM): A WDFT-Based Anthropomorphic Noise-Robust Feature Extraction Algorithm for Speech Recognition" Proc. of Interspeech'2005 (Eurospeech), pp. 713 -716, September, 4-8, 2005, Lisbon, Portugal

[9] Malina, R.M. and Bouchard, C. and Bar-Or, O., "Growth, Maturation, and Physical Activity", 2nd ed., Human Kinetics, 2003, 701 p.-

[10] W. Tecumeseh Fitch and J. Giedd, "Morphology and Development of the Human Vocal Tract: a Study Using Magnetic Resonance Imaging", J. of Acoust. Soc. Am., 106 (3), Pt. 1, September 1999, pp. 1511 - 1522

[11] Bone, D. and Black, M.P. and Li, M. and Metallinou, A. and Lee, S. and Narayanan, S. "Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors", Proc of Interpeech'2011, 3217-3220.

[12] R. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," in Mach. Learn., 2000, pp. 135 -168.