ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Modulation Spectrum Analysis for Speaker Personality Trait Recognition

Alexei V. Ivanov, Xin Chen

Knowledge Technologies, Pearson
4040 Campbell Ave., Suite 200, Menlo Park, California 94025, USA
alexei_v_ivanov@ieee.org, xin.chen@pearson.com

Abstract

We explore the utility of individually selected modulation spectral features for speech and speaker characterization in general, and specifically to prediction of the perceived speaker personality profile. We suggest a method of construction of a sparse feature space and a method of finding the approximately best feature subset for attributing a specific characteristic of speech or speaker. The current selection method is based on the Kolmogorov-Smirnov statistical test applied to individual features. We assume that the characterization task is defined empirically and no a-priory theory exist to explain characteristic attribution processes. Experimental results indicate that employment of selected modulation spectral features works better than the current state-of-the-art in prediction of personality traits.

Index Terms: speech characterization, modulation spectrum analysis, feature selection

10.21437/Interspeech.2012-92

1. Introduction

Speech and speaker characterization is concerned with attribution of a particular characteristic to a speech sample originated by a speaker in an objective and consistent manner. Spoken communication is a more capacious channel in comparison with textual. Meaning attribution of the natural communication act can be aided by determining and interpreting the paralinguistic aspects of the message. Spoken language, as an aspect of human behaviour, can also be used as an information source for acquisition of psychometric information in general and personality trait recognition in particular.

Recent work [1–3] has put forward a tentative benchmark in this field of study. However the data used was coming from a single actor and the lexical content was fixed throughout the whole experiment. A study of self-report personality trait recognition from speech in a more realistic dialogue environment was given in [4]. It has been found that personality traits, which are determined from spoken content in the most reliable manner, are conscientiousness and extroversion. Although self-report personality traits are not necessarily required to coincide with those, judged by a third party observer, the baseline performance level of the current perceived Personality Trait Recognition Challenge [5] confirms this finding of [4].

The Big Five model [6] is a generally used personality model. It describes human personality as a vector of five values corresponding to bipolar traits:

- Openness to experience: A preference to a varying experience, an appreciation for art, emotion, adventure, etc.
- Conscientiousness: A tendency to have a planned behaviour (as opposed to spontaneous responses), a manifestation of self-discipline.
- Extroversion: "Energetic" behaviour, an outgoing attitude, seeking the company of others.
- Agreeableness: Compassion and cooperativeness (as opposed to suspicion).
- Neuroticism: A tendency to "mood swings", a tendency to negative emotions such as anger or vulnerability.

Modulation spectrum estimation has previously been tried with speech detection [7–9] and speech [10] and speaker [11] recognition. Apart from its applicability to speech detection this approach suffers from sparseness of the signal representation and special precautions shall be taken in order to compress feature space. In this paper we explore a combination of the modulation spectrum analysis (MSA) for feature extraction and Kolmogorov-Smirnov statistical test (KST) for feature selection in the task of speech characterization and specifically perceived speaker personality trait recognition.

The paper is organized as follows: section 2 discusses modulation spectrum analysis for feature extraction purposes; section 3 is devoted to the description of feature selection process; section 4 describes the implemented classification system and the personality trait recognition experiments.

2. MSA Feature Extraction

The MSA method is outlined in Figure 2. The method uses a temporal sequence of short-time Fourier transform (STFT) spectral representations of the original signal as its input. Each of the spectral bins is considered a signal in time, for which another spectrum is obtained. This

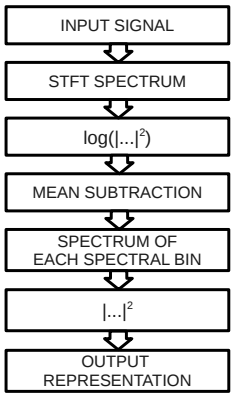

Figure 1: *Schematic structure of the MSA feature extraction.*

spectrum reflects how fast the energy of the respective frequency band is getting alternated through time.

The transformation in the log-spectral domain is even more revealing. By taking a spectral representation of the individual log-spectral amplitude (power) trajectories it is possible to isolate and characterize the dynamical properties of the speech production apparatus, abstracted from the properties of the excitation source. Mean subtraction in the log-spectral domain can eliminate most of the excitation source component of the signal. The resulting signal representation is three-dimensional, having frequency, modulation frequency and time as axes.

The output is approximately invariant with respect to the incoming signal scale; hardly varying when the input signal is enhanced by as much as 20 dB. Thus, the dynamic range of the method is reasonably large compared to the dynamic range of the pulse-coded modulation (PCM) digital signal representation.

Each of the spectral transformations might have a different analysis interval duration. In order to capture the speech source dynamics in the most complete manner, the output is computed for each of the reasonable combinations of the analysis interval duration. The final output of the MSA method is a family of three-dimensional streams of features.

3. KST Feature Selection

The output of the MSA method is not directly suitable for use as a feature stream in speech characterization because of the large number of features. A separate procedure needs to select useful features for the particular speech characterization task.

The speech characterization task is defined statistically as a representative collection of speech samples that are known to have different quantitative characteristics along the chosen qualitative dimension. Thus, the feature distributions, conditioned on different quantitative characteristics, are defined in an empirical manner.

Statistical selection of a set of useful features is generally a complex task. In the case of large-dimensional feature space it requires a prohibitively large amount of supporting data. Consequently the MSA-based method of speech and speaker characterization must rely on a kind of engineering approximation to derive an estimate of the useful feature set. One such possibility is to evaluate each feature independently of the rest with the help of KST. This method may be applied either to individual features themselves or, in order to reduce the computational complexity, to statistics, estimated over that feature (e.g. statistical moments of the feature distribution within a single speech sample). Application of KST for the feature selection process is discussed in greater detail in [12].

In general KST aims to reject at the specified level of significance p the null-hypothesis H_0 that two random variables have identical distributions. For a given pair of random variables X and Y the Kolmogorov-Smirnov statistics $D_{X,Y}$ is the largest observed discrepancy between the estimated cumulative distributions across the sample space (i.e. $\forall z \in (-\infty, \infty)$):

$$D_{X,Y} = \sup_{\forall z \in (-\infty,\infty)} |\hat{P}(X \leq z) - \hat{P}(Y \leq z)|. \quad (1)$$

The usefulness of KST in application to feature selection for speech and speaker characterization results from the absence of the explicit analytical assumptions on the form of the conditional feature distributions. It is possible to estimate a probability that differently conditioned feature distributions are identical even in the case when these distributions are defined empirically.

KST is only applicable to the uni-variate case, when

$$P(X \leq z) = 1 - P(X > z) \quad (2)$$

and, thus, according to (1), $D_{X,Y}$ is invariant in respect to orientation of the sample space. In this case there is no difference whether we traverse it from $-\infty$ to ∞ or the other way around. Substitution of the '$\leq$' sign with a '$<$' sign during order inversion is not essential for the definition of the Kolmogorov-Smirnov statistics.

Unfortunately, because the invariance (2) does not hold for multi-variate distributions, there is no straightforward generalization of KST to multivariate analysis. See [13] for more detailed discussion on existing multivariate generalizations,

In practice the feature selection process might be implemented either as a standard statistical hypothesis rejection at the predefined significance level, or, alternatively, a selection of a predefined number of features having the smallest associated probability.

4. Experiments

4.1. Feature Extraction and Selection

The experimental results are obtained while working with the data provided by the organizers of Interspeech 2012

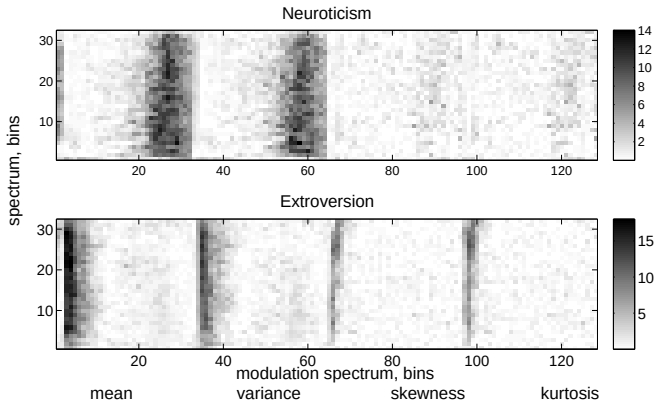

Figure 2: *Inverse log probability of $P(x|0) = P(x|1)$ of statistical moments of MSA features. Comparison between "Neuroticism" and "Extroversion" personality traits. FFT size is 64 points.*

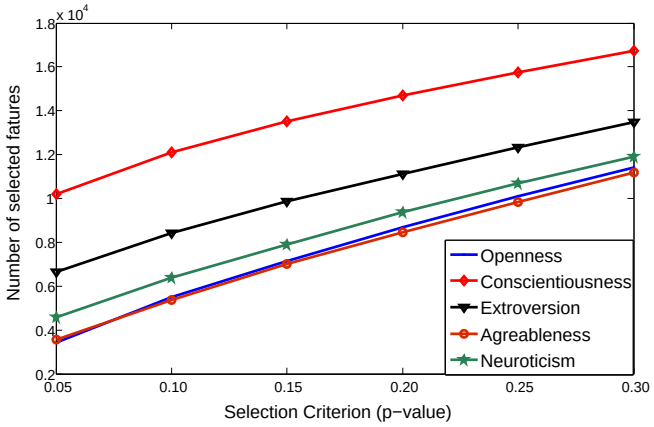

Figure 3: *Rate of decay in feature content over a range of KST p-values for different perceived psychological traits.*

Speaker Trait Challenge [5]. There is no alternation in the definitions of data splitting or labelling.

Four MSA feature streams are computed. Each is configured to have equal FFT sizes for both spectrum calculations in MSA. The FFT size ranges from 16 to 128 points. Each of the spectral bins in the two-dimensional array is represented by four statistical moments (mean, variance, skewness and kurtosis) of its distribution inside a specific utterance. Thus, the total size of MSA feature vector before selection is equal to 21760 values. The total feature count after application of the four statistical functionals is $4 \times (8^2 + 16^2 + 32^2 + 64^2)$, as only a half of FFT values along each dimension is useful. Features from the baseline system (6125 values per speech sample) are also added to a common raw pool before feature selection.

Selection of features is done with two criteria: dissimilarity of the feature distributions, conditioned with different class labels (e.g. 'neurotic' vs 'non-neurotic' speech) in the training data; similarity of feature distributions over training and development data. The rationale behind the second criterion is to avoid working with features, which happen to violate representativeness of the training set.

Figure 2 presents an example of KST-based feature evaluation. This example is given for two personality traits, namely "Neuroticism" and "Extroversion". Four squares, corresponding to the statistical moments (mean, variance, skewness and kurtosis) of the MSA features, are placed horizontally adjacent to each other. The spectral range is given along the Y-axis and modulation spectrum runs along the X-axis. A variable shade of grey is used to reflect the inverse log probability of the fact that the distribution of that particular feature for the positive trait label ("Neurotic" and "Extrovert" respectively) is the same as that for the negative trait label ("Non neurotic" and "Introvert" respectively). The analysis is done with the whole set of labelled data of the Challenge. More useful features are coloured darker.

As can be seen, "Neuroticism" and "Extroversion" traits have very distinct patterns of useful features. They are spatially localised for both classes, which is important if one considers creation of parametric feature selection models for recognition. Apparently, attribution of the "Neurotic" label has something to do with abrupt alternation of the input signal especially in the higher spectral range. While the "Extroversion" perceived trait is linked with differences in speech pace in the lower modulation-spectral range across the whole spectral frequency range.

A comparative example of the dependency of feature survival rate over a range of KST p-values for different perceived psychological traits is given in Fig. 3. The un-pruned feature stream contains all possible spectro–modulation-spectral resolutions together with the baseline set of features. It is evident that the rate of decay in feature content is nearly constant and equal across all perceived psychological traits, although different traits have different numbers of surviving features at a given KST p-value. Compare for example a number of surviving features for "Neuroticism" and "Extroversion". In agreement with the picture, presented in Fig. 2, "Extroversion" trait has a larger number of useful features at any given p-value.

4.2. Recognition Experiment

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) [14]. Specifically, an open source implementation "icsiboost" is used. Training is done for 6000 iterations. We also use the idea of combining multiple weak classifiers obtained by training on randomly sub-sampled training set.

A search for an optimal operating point has been performed through a reasonably wide range of p-values for each of the selection criteria. Generally if the p-values are chosen too lax the amount of surviving features is going to be overwhelmingly large causing slow convergence and possibilities to get a suboptimal final solution and have a model lacking the generalisation power. Alternatively if the the feature pruning is too aggressive, than valid information is going to be eliminated, which also results in the performance degradation. The optimal values depend on the particular speaker trait. Table 1 presents a summary of the optimal parameter values.

Table 1: *Optimal parameters for feature selection and classification. "P 0_1" – p-value for selection over dissimilarity of the feature distributions having different class labels;"P t_d" – p-value for selection over similarity of the feature distributions of training and development data; "NFeat" – total number of features surviving the selection; "NFeat BL" – number of surviving features from the Challenge Baseline feature set.*

Personality Trait	P 0_1	P t_d	NFeat	NFeat BL
Openness	0.05	1	6719	783
Conscientiousness	0.25	0.5	6904	1487
Extroversion	0.2	1	13425	3148
Agreeableness	1	0.6	9373	1765
Neuroticism	0.2	1	9970	2375

Table 2: *Performance of predictors of individual Big Five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism),"CORR" – number of correctly labelled utterances. "UA" – unweighed average recall in percent,"Acc" – Accuracy (weighted average recall) in percent,"p-value" – probability to see at least the observed number of correct recognitions assuming that the recogniser is no different from the baseline,"MSA" – The best accuracy of the slected MSA-only features, "MSA+BL" – The best accuracy of the pruned joint MSA & Baseline pool*

Set	Test				Development	
Trait	CORR	UA	Acc	p-value	MSA	MSA+BL
O	121	56.353	60.199	0.4162	75.956	74.863
C	155	77.035	77.114	0.8347	72.131	75.410
E	154	76.486	76.617	0.4704	85.792	85.792
A	135	**67.232**	67.164	0.0175	70.492	72.678
N	138	**69.204**	68.657	0.1694	76.503	75.956
Mean	140.6	**69.262**	69.950	0.3374	76.175	76.940

Table 2 summarizes recognition results on the official Speaker Personality Challenge evaluation set. Recognition accuracy for all but one trait is better than the baseline. However, most of the improvements are not significant with the amount of test data provided. More test data is needed in order to ultimately establish the degree of improvement offered by the proposed method. Statistical significance of the accuracy difference is estimated with a one-tail binomial test. P-value is estimated as a probability of seeing at least the observed number of successful recognitions under the null-hypothesis that the baseline accuracy is a valid maximum likelihood estimate of the probability to make a correct recognition.

5. Conclusions

We have found that application of the modulation spectrum–based feature extraction technique in combination with individual feature selection by Kolmogorov-Smirnov statistical test is advantageous in the task of speech characterization. Specifically in application to prediction of the perceived speaker personality profile it allows to obtain results better than the current state of the art baseline for many personality traits.

We have discovered that the features that survive the selection process exhibit good spacial localization in the modulation-spectral domain, which potentially permits construction of the feature selectors based on parametric statistical modelling. Different personality traits apparently have different useful features.

6. References

[1] T. Polzehl, S. Moeller, and F. Metze, "Automatically assessing acoustic manifestations of personality in speech," in *Spoken Language Technology Workshop, 2010 IEEE*, 2010, pp. 7 –12.

[2] ——, "Automatically assessing personality from speech," in *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, ser. ICSC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 134–140.

[3] ——, "Modeling speaker personality using voice," in *Proc. Interspeech 2011, pp. 2369-2372, Florence, Italy*, 2011.

[4] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Proc. Interspeech 2011, pp. 1549-1552, Florence, Italy*, 2011.

[5] B. Schuller, S. Steidl, A. Batliner, E. Noeth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, G. Bocklet, T. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *Proc. Interspeech 2012, ISCA, Portland, OR, USA*, 2012.

[6] J. Digman, "Personality structure: emergence of the five factor model," *Annual Review of Psychology*, vol. 41, pp. 417–40, 1990.

[7] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of Speech from Nonspeech Based on Multiscale Spectrotemporal Modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 920–930, May 2006.

[8] J.-H. Bach, B. Kollmeier, and J. Anemüller, "Modulation–Based Detection of Speech in Real Background Noise: Generalization to Novel Background Classes," in *Proc. of Int. Conf. on Acoust. Speech and Signal Processing (ICASSP)*, March 2010, pp. 41–44.

[9] A. V. Ivanov and G. Riccardi, "Automatic turn segmentation in spoken conversations," in *Proc. of Interspeech'2010, Makuhari, Japan*, 2010.

[10] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 3, apr 1997, pp. 1647 –1650 vol.3.

[11] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[12] A. V. Ivanov and G. Riccardi, "Kolmogorov–Smirnov test for feature selection in emotion recognition from speech," in *Proc. of ICASSP'2012, Kioto, Japan*, 2012.

[13] R. Lopes, I. Reid, and P. Hobson, "The two-dimensional Kolmogorov-Smirnov test," in *Proc. XI Int. Workshop on Adv. Computing and Analysis Tech. in Physics Res.*, April 2007.

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