Multi-System Fusion of Extended Context Prosodic and Cepstral Features for Paralinguistic Speaker Trait Classification

Michelle Hewlett Sanchez, Aaron Lawson, Dimitra Vergyri, Harry Bratt

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.

{mhewlett,aaron,dverg,harry}@speech.sri.com

Abstract

As automatic speech processing has matured, research attention has expanded to paralinguistic speech problems that aim to detect beyond-the-words information. This paper focuses on the identification of seven speaker trait categories from the Interspeech Speaker Trait Challenge: likeability, intelligibility, openness, conscientiousness, extraversion, agreeableness, and neuroticism. Our approach combines multiple features including prosodic, cepstral, shifted-delta cepstral, and a reduced set of the OpenSMILE features. Our classification approaches included GMM-UBM, eigenchannel, support vector machines, and distance based classifiers. Optimized feature reduction and logistic regression-based score calibration and fusion led to results that perform competitively against the challenge baseline in all categories.

Index Terms: speaker traits, prosody, MFCCs, Gaussian mixture modeling

1. Introduction

As automatic speech processing has matured, research attention has expanded to applications focusing on paralinguistic speech problems that aim to detect this "beyond-the-words" information. Researchers have focused on automatically deriving speaker characteristics from speech and classifying speakers into categories ranging from identity [1], age [2], language [3], dialect [4], idiolect and sociolect [5] to truthfulness [6], psychological health [7], personality traits [8], and emotion [9]. This paper focuses on the identification of seven speaker trait categories from the Interspeech Speaker Trait Challenge: likeability, intelligibility, openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Prosodic features in speech, such as speaking rate, pitch, energy or intensity, and pause duration have been used for other paralinguistic problems in previous work [10, 11, 9]. In addition, other acoustic features such as voice quality [11], spectral features [10, 9], and mel frequency cepstral coefficients (MFCCs) [9] have also been explored.

This paper is organized as follows. In Section 2, concepts that have been found successful in other speech areas like speaker identification are applied to speaker trait tasks. Score fusion on these systems is also explored for added improvement. Section 3 describes the results for each type of classification. Finally, Section 4 gives an overall summary of this work.

2. Features and Classifiers

This section discusses both the features and the classifiers that were used for the speaker trait challenge tasks.

All the tasks described in Section 3 are binary classifications. Some of the methods discussed in this section require a large amount of training data, while the available data tagged with paralinguistic labels are limited. We show that we can successfully use other available English data, labeled only with speaker information, to train some of our model parameters.

2.1. Features

As in past literature for these tasks, acoustic features such as prosodic and spectral characteristics are modeled. The prosodic features are extracted using Algemy, which is software designed at SRI for the development and extraction of prosodic features.

In addition to the given 6125 features [12], we used four types of features, two based on prosodic measurements and two based on spectral measurements.

2.1.1. Basic Prosodic Features

In recent years, a large amount of work has been done on prosodic features for a variety of speech tasks. Prosodic features have been proven very useful in previous paralinguistic work as seen in Section 1. The prosodic features include several statistics on:

- Pitch, extracted as the fundamental frequency (*F0*)
- Pitch normalized by subtracting the overall mean of the speaker (*F0Norm*)
- Pitch normalized by subtracting the overall mean of the speaker and dividing by the overall standard deviation of the speaker (*F0ZNorm*)
- Voiced energy: the root mean square of the amplitude of the signal over only the voiced regions or regions with speech, normalized by dividing by the maximum energy during the session (*EnergyOverVoiced*)
- Spectral Tilt, defined as the slope of the line that connects the values of the formants, extracted using Praat [13] (*SpecTilt*)

The pitch and energy signals were extracted using a pitch tracking algorithm implemented by the *get_f0* function [14]. This pitch tracking algorithm estimates, every 10 milliseconds (or every frame), the fundamental frequency, the probability that the frame contains voiced speech, and the root mean square of the amplitude of the speech signal.

All pitch and energy features are converted to a log scale. For the pitch features, two different normalizations were tried as mentioned above: (1) normalization of each speaker's pitch by the overall speaker mean *F0Norm* or (2) normalization by the overall speaker mean and standard deviation *F0ZNorm*.

For the energy features, normalization is always necessary since the energy of the speech signal is strongly correlated with aspects that should be removed, like the channel characteristics and the distance of the speaker from the microphone. Since very low energy regions of the signal typically correspond to

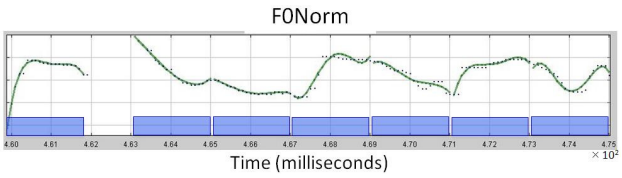

Figure 1: Example of the prosodic polynomial coefficients of order 5 with non-overlapping windows for *F0Norm* as a function of time.

pauses or silence, which is strongly affected by the background environment, we chose to normalize the energy features by dividing by the maximum energy value of the session. This value is not affected by the low energy regions (as the mean energy of the session would) and, hence, is not affected as much by the background noise present in the signal.

The statistics we extract on the prosodic signals are obtained over the whole utterance. These statistics consist of the maximum, the minimum, the range (the maximum minus the minimum), the mean, and standard deviation. In the experiments that use the basic features, one set of pitch features from *F0*, *F0Norm*, and *F0ZNorm* and one of either the energy or spectral tilt features from *EnergyOverVoiced* and *SpecTilt* are used, giving 10 features in total. These features were used because of their usefulness in past research [15].

2.1.2. Prosodic Polynomial Coefficients

The second type of prosodic features are polynomial coefficients of the prosodic features. To obtain the polynomial coefficient features, the Legendre polynomial regression of order five is calculated on one set of pitch features and one set of either the energy or spectral tilt features giving 12 features in total. These prosodic polynomial coefficient features contain information about the change in prosody values in small overlapping regions. The dynamics of a person's speech can be very subtle; therefore, using this technique allows thorough and accurate modeling of these intricacies in the speech. This technique of using uniform regions where the features in each region are order-five polynomial coefficients is a method recently used in speaker identification [16].

In Figure 1, we plot the normalized pitch values (the blue dots) as a function of time. The uniform regions have length 200 milliseconds with a shift of length 100 milliseconds. The blue rectangles show the non-overlapping uniform regions, every 200 milliseconds in this case. The polynomials are the green curves, one for each uniform region. Pitch features only exist in voiced regions, so the polynomial regression interpolates two fill in pitch values that do not exist.

2.1.3. Spectral Features

We use as spectral features the mel frequency cepstral coefficients (MFCCs). These features are extracted over a window length of 25 milliseconds that is shifted every 10 milliseconds to have more features for modeling the waveform. The spectral features used in Section 3 are based on 13 coefficients with cepstral mean subtraction in addition to the first, second, and third derivatives, resulting in a feature vector size of 52.

2.1.4. Shifted Delta Cepstrum (SDC)

MFCCs were extracted to an original dimensionality of 7 coefficients (0-6). These were then transformed using the shifted-delta-ceptrum technique [17] for increasing feature context using the typical 7-1-3-7 shifting and stacking parameters to yield a final feature vector of 56 (49 SDC+7 cepstral) dimensions.

2.2. Classifiers

2.2.1. Support Vector Machines

SVM Light with default parameters was used in all SVM linear classification experiments [18]. Normalization was performed on the features so no one feature affects the objective function of the SVM more than any other. To this end, the features used in the training data are normalized to mean zero and variance one. The statistics of each feature, which are found from the training data, are then applied to the test data to normalize based on the same criteria.

2.2.2. UBM-GMM MAP System

In this system, a Gaussian mixture model (GMM) is trained with data from many speakers speaking English (since we did not have enough German, Dutch, or French data) with various recording conditions to create a universal background model (UBM).

For a GMM classifier, we need to train one GMM for each target class. All of the speaker trait challenge tasks are binary tasks, which means there are two classes, A and B. The UBM is adapted to the data for each class using a maximum a posteriori (MAP) estimation of the means [19]. The result is two GMMs, one for class A and one for class B.

2.2.3. UBM-GMM Nuisance Compensation System

Nuisance compensation (NC) has been proven useful in both speaker and language identification. It is normally used to account for effects due to speech recorded on different channels, with different styles, or with different background noise. For these speaker trait challenge tasks, the nuisances are these channel effects in addition to the speaker effects. These speaker-dependent effects should be reduced so that the trait differences in the speakers can be detected in a speaker-independent way.

As in the GMM-UBM MAP system, the UBM is also trained on held-out data for the GMM-UBM NC system. The difference in the NC case is that each class' model, which is trained using the data from each class, is adapted assuming the mean is given by $\mu = \mathbf{m} + \mathbf{U}\mathbf{x}$, where $\mathbf{m}$ is the concatenated means of the UBM, $\mathbf{U}$ is a low rank matrix, and the columns are the directions of channel and speaker variability called *eigenchannels*, and $\mathbf{x}$ is learned from the sample [20].

Each class model's mean is estimated on the training data for each class. The eigenchannel matrix is trained using the data for both classes. The number of eigenchannels chosen varies, but cannot exceed the number of sessions available for training, which for these speaker trait tasks is the number of speakers in the training data.

In both GMM-UBM MAP and GMM-UBM NC, the overall systems are the same. After the two GMMs are created, the test waveform is used to extract features to obtain a feature sequence. These features are the input to each GMM and a score for each class is obtained.

The two scores are the log likelihoods computed from each GMM. Since both GMMs were created using the same set of training data, we directly compared the two scores to determine which class the test segment belonged to.

2.2.4. Eigenchannel SDC System

This system is based on a context modeling approach popular for language identification. Modeling begins with a 2048 component UBM built from a very large amount of data covering most of the languages in the NIST Language Recognition Evaluation corpus. An eigenchannel compensation system [21] was

trained using the training data for each sub-task, with a co-rank of 25 geared towards removing speaker variability factors.

2.2.5. Feature Reduction System

Due to the very large number of features provided with the challenge (6125 dimensions)[12], feature reduction was important. Use of the BOOST algorithm [22] did not produce reliable results with this feature set, probably due to the huge amount of noise in the features. A simpler approach was devised that measured the efficacy of each individual feature on its own. This was done by treating each feature component as a classifier with which the training data was evaluated, using maximum average recall as the optimization measure. Certain features are positively correlated with a given class and others negatively correlated, though the vast majority were not strongly correlated. Those features with the highest average recall as individual classifiers were selected, for a total of 75 positively correlated and 75 negatively correlated features per task. Results using the reduced feature set in a simple Euclidean Distance based (L2) classifier generally outperformed the SVM using all features, as did the SVM using the reduced set.

2.2.6. Fusion and Calibration

Logistic regression was used to calibrate the resulting scores from the various systems and to fuse the scores to create proper likelihoods. This greatly facilitated the fusion of results from disparate sources where scores may be on very different scales and cover very different ranges. Fusion is used when systems (A and B) are added together like A+B.

2.3. Systems

Experiments are performed using seven different types of systems described below. If the system has various parameters, experiments were performed on every possible setting.

2.3.1. SVM Challenge System

The SVM challenge system uses the given feature vector of size 6125 described in [12] with SVM modeling.

2.3.2. SVM Prosodic System

The SVM prosodic system uses the basic prosodic feature vector of size 10 with SVM modeling.

2.3.3. MAP Prosodic Polynomial Coefficient System

This system uses the prosodic polynomial coefficient features of size 12 with the MAP UBM-GMM system. The size of the UBM was varied on a log scale from 32 to 1024 Gaussians.

2.3.4. NC Cepstral System

This system uses the 52 dimensional cepstral features with the NC UBM-GMM system. As in the previous case, the size of the UBM was varied on a log scale from 32 to 1024 Gaussians. Another parameter of the NC cepstral system is the number of eigenchannels. The number of eigenchannels was varied choosing 10, 30 or 50. 10 was the optimal in most cases.

2.3.5. Eigenchannel (EC) MFCC SDC

This system is the eigenchannel SDC system with the shifted delta cepstrum features.

2.3.6. L2 Feature Reduction (Feat Red)

This system uses the top 75 positively correlated (pos) features or top 75 negatively correlated (neg) features from the 6125 features and the L2 classifier.

System	Likeability	Intelligiblity
Chall. SVM Baseline	58.5%	61.4%
Chall. RF Baseline	57.6%	65.1%
SVM Challenge	56.8%	58.8%
SVM Prosodic	57.1%	54.8%
Pros Poly Coeff	61.5%	58.8%
NC Cepstral	58.3%	64.1%
EC MFCC SDC	64.6%	61.5%
L2 Feat Red (pos)	58.3%	60.5%
L2 Feat Red (neg)	52.6%	58.7%
SVM Feat Red (pos)	50.4%	60.2%
SVM Feat Red (neg)	52.5%	64.6%
Fusion*	**66.0%**	**67.3%**

Table 1: Unweighted Average (UA) results on the development data for the Likeability and Intelligibility Speaker Trait Challenge. Fusion* is Pros Poly+NC Cep+SDC+L2 (pos)+L2 (neg)+SVM (neg).

2.3.7. SVM Feature Reduction (Feat Red)

This system uses the top 75 positively correlated (pos) features or top 75 negatively correlated (neg) features from the 6125 features and the SVM classifier.

3. Experiments

3.1. Experimental Setup

Unweighted average or UA $= (\text{recall}(A) + \text{recall}(B))/2$ is the evaluation measure for the challenge. In the results in Section 3.3, the challenge (chall.) baseline performance is compared to the ten different experimental setups that were tried.

3.2. Databases

The databases for the speaker trait challenge were described in [12].

3.3. Results and Analysis

Table 1 shows the results for the likeability and intelligibility sub-challenges on the development data. The fusion of six systems outperformed any system individually. The fused system also outperforms the SVM challenge baseline by 7.5% for likeability and 6% for intelligibility.

Table 2 shows the results for the personality sub-challenge on the development data. The fusion of six systems outperforms any system individually. The fused system also outperforms the SVM challenge baseline by 9% for openness, 5% for extraversion, and 3% for neuroticism. For the two remaining categories, conscientiousness and agreeableness, the fused system performed similarly to the baseline.

Table 3 shows the results for all sub-challenges on the test data. The fused systems were the only ones that were submitted for testing.

4. Summary

We explore techniques and features that have been proven useful in other speech tasks like speaker and language identification in addition to the features given in the challenge. Our fused system beats the challenge baseline in six of the seven categories on the development data. Our fused system also beats the challenge baseline in openness on the test data and performs competitively on intelligibility. We believe there may be a mismatch between the development and test data because after seeing our fused test results, we tested a single system that performed well in development and it did not perform well on the test set.

System	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Chall. SVM Baseline	60.4%	74.5%	80.9%	67.6%	68.0%
Chall. RF Baseline	57.7%	74.9%	82.8%	67.2%	68.9%
SVM Challenge	54.6%	72.1%	83.1%	65.8%	69.7%
SVM Prosodic	50.0%	64.3%	78.1%	61.4%	65.0%
Pros Poly Coeff	59.6%	69.6%	73.2%	62.7%	67.8%
NC Cepstral	57.2%	63.9%	73.2%	63.2%	66.4%
EC MFCC SDC	60.3%	70.0%	77.0%	55.7%	56.8%
L2 Feat Red (pos)	68.6%	71.5%	76.5%	66.4%	61.7%
L2 Feat Red (neg)	61.7%	73.5%	79.2%	63.3%	63.5%
SVM Feat Red (pos)	62.3%	71.6%	78.1%	58.9%	66.6%
SVM Feat Red (neg)	58.4%	72.8%	82.0%	64.2%	66.5%
Fusion*	**69.3%**	74.4%	**85.8%**	**67.7%**	**71.4%**

Table 2: Unweighted Average (UA) results on the development data for the Personality Speaker Trait Challenge. Fusion* is Pros Poly+NC Cep+SDC+L2 (pos)+L2 (neg)+SVM (neg).

Data	Fusion Features	UA
Like.	Pros Poly+SDC+L2 (pos)	58.2%
Intel.	Fusion*	69.1%
Open.	Pros Poly+NC Cep+SDC+L2 (neg)+SVM (neg)	**62.5%**
Cons.	Fusion*	78.6%
Extr.	Fusion*	71.4%
Agre.	Fusion*	59.9%
Neur.	Fusion*	60.1%

Table 3: Unweighted Average (UA) results on the test data for all categories of the Speaker Trait Challenge. Fusion* is Pros Poly+NC Cep+SDC+L2 (pos)+L2 (neg)+SVM (neg).

5. Acknowledgements

This material is based upon work partially supported by the U.S. Army Research Office under Contract No. W911NF-12-C-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Army Research Office.

6. References

[1] E. Shriberg, "Higher-Level Features in Speaker Recognition," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed., no. 4343 in *Lecture Notes on Artifical Intelligence*. Springer, 2007, pp. 241–259.

[2] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on Acoustic Speech Signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975–1985, September 2011.

[3] A. Stolcke, M. Akbacak, L. Ferrer, S. Kajarekar, C. Richey, N. Scheffer, and E. Shriberg, "Improving Language Recognition with Multilingual Phone Recognition and Speaker Adaptation Transforms," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 256–262.

[4] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, and A. Mandal, "Effective Arabic Dialect Classification Using Diverse Phonotactic Models," in *Interspeech*, Florence, Italy, 2011.

[5] T. Schultz, "Speaker Characteristics," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed., no. 4343 in *Lecture Notes on Artificial Intelligence*, 2007, pp. 47–74.

[6] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Pittsburgh, PA, USA, 2006.

[7] H. K. Keskinpala, T. Yingthawornsuk, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, "Screening for High Risk Suicidal States using Mel-Cepstral Coefficients and Energy in Frequency Bands," in *European Signal Processing Conference*, Poznan, Poland, 2007, pp. 2229–2233.

[8] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Frank, "Recognition of Personality Traits from Human Spoken Conversations," in *Interspeech*, Florence, Italy, 2011.

[9] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, 2007.

[10] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog," in *International Conference on Spoken Language Processing*, Denver, CO, USA, 2002.

[11] M. Lugger and B. Yang, "The Relevance of Voice Quality Features in Speaker Independent Emotion Recognition," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, 2007.

[12] B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Viciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Interspeech*, Portland, Oregon, 2012.

[13] "Praat: Doing phonetics by computer," *http://www.fon.hum.uva.nl/praat/*.

[14] D. Talkin, *Robust Algorithm for Pitch Tracking*. Elsevier Science, 1995.

[15] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, and W. Jarrold, "Using Prosodic and Spectral Features in Detecting Depression in Elderly Males," in *Interspeech*, Florence, Italy, 2011.

[16] M. Kockman, L. Burget, and J. Cernocky, "Investigations into Prosodic Syllable Contour Features for Speaker Recognition," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, 2010, pp. 4418–4421.

[17] B. Bielefeld, "Language Identification using Shifted Delta Cepstrum," in *Fourteenth Annual Speech Research Symposium*, 1994.

[18] "Svmlight support vector machine toolkit," *http://svmlight.joachims.org*.

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

[20] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[21] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP Estimators for Speaker Recognition," in *Eurospeech*, Geneva, Switzerland, 2003.

[22] R. Schapire, "The Strength of Weak Learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.