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

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

Pitch and Intonation Contribution to Speakers' Traits Classification

Claude Montacié [1], *Marie-José Caraty* [2]

[1] Laboratory, Paris Sorbonne University, 28 rue Serpente, 75006, Paris, France
[2] LIPADE laboratory, Paris Descartes University, 45 rue des Saints-Pères 75006, Paris, France
Claude.Montacie@paris-sorbonne.fr, Marie-Jose.Caraty@ParisDescartes.fr

Abstract

The article describes the system we submitted for the three sub-challenges of INTERSPEECH 2012 Speaker Trait Challenge for the classification of the five personality traits of OCEAN, likability and intelligibility. The system was based on a two-class SVM-classifier using leave-one-speaker-out cross-validation to optimize SVM complexity parameter and to select the feature set for trait classification. Pitch, intonation and spectrum contributions to speakers' traits classification were studied. One feature set was specially designed for intonation modeling. Variations from the official feature set based on pitch and spectrum were assessed in traits classification. Preliminary results on the Test set have shown a significant improvement of 4.7% on the likability trait, an improvement of 1.1% on the intelligibility trait and no improvement on the OCEAN traits compared to the official baseline unweighted accuracy results.

Index Terms: speaker's trait detection, pitch, intonation

1. Introduction

INTERSPEECH 2012 Speaker Trait Challenge [1] aims at classifying speakers' traits in two classes. The first sub-challenge concerns the classification of the five personality traits from speech clips of 10 seconds mean duration: *Openness* (O), *Conscientiousness* (C), *Extraversion* (E), *Agreeableness* (A), *Neuroticism* (N). The Five-Factor Model of Personality [2] was the result of studies on lexical analysis and etymology of adjectives terms related to personality which are widely accepted [3]. The second sub-challenge concerns the *Likability* trait (L) of the speaker's voice using speech clips of 3 seconds mean duration [4]. The third sub-challenge concerns the *Intelligibility* trait (I) for pathologic voices using speech clips of 3 seconds mean duration. This trait is also studied in order to assess the quality in speech coding or enhancement [5]. The transcription of the corpora into traits is performed from listening by ratings. In the challenge only the non-verbal cues are considered for the transcription.

A reliable detection of speakers traits/states lets envisage a multitude of applications. In [6], a taxonomy of individual's traits and states was proposed according to temporal scale and 14 broad scopes of applications are considered in the domain of vocal human-machine interaction. Sociological environment pushes to develop other traits of personality reflecting cognitive differentiation such as persuasiveness, social attractiveness, etc. In [7], an example is found in the review on voice attributes of persuasiveness of telemarketers' voice. In the article and tied to vocal behavior, recommendations for operators are reported. Five of the ten recommendations concern prosodic characteristics: -speaking rate of 150-200 words per minute (wpm), -rate slightly faster (no superior to 40 wpm) than listener, -speak fluently with few pauses, unnatural hesitations, or disfluencies, -greater pitch, volume and stress/emphasis variety, -volume moderately loud, -slow adjustment to converge rate and other cues. From research on identification of vocal cues in non-verbal paralanguage, knowledge and comprehension could be expected to define "how to have a persuasive voice".

In many studies, prosodic characteristics have been found linked to personality traits [8]. In [9], alterations of voice recordings have shown that listeners use acoustic cues such as pitch level, speech rate and loudness in making personal judgment of the speaker. In synthesis experiments, pitch level, pitch range articulation rate and loudness were shown significant in rating five personality traits other than OCEAN traits [10]. Prosody includes pitch, intonation, stress, loudness, rhythmic, speech rates, pauses, etc. Characterized by a pitch level and contour types, intonation contours arouse sensations in the listener and were shown to have a major influence in perception of emotions and mood [11, 12].

For the challenge, we paid a particular attention on the acoustic cues that should impact classification performance according to related works on personality traits and emotions. At studying the official audio feature set of the challenge, we looked for improvement in the representation of features tied to pitch. In this purpose, -additional functionals have been first applied to pitch representation; then, intonation contours have been stylized by the INTSINT targets [13] and functionals were developed to extract features from intonation targets. In order to assess performance improvement and guide our choices, we chose as criterion the leave-one-speaker-out cross-validation (LOSO-cv). This criterion was used for the various choices like SVM complexity parameter and feature set selection.

The paper is organized as follows. In section 2, the Baseline System (BS) is described and its LOSO-cv results on the three sub-challenges are given. In section 3, the official acoustic feature set related to the pitch information is extended, the intonation model and corresponding features are described. The contribution of pitch and intonation features to speakers' traits classification is studied. In section 4, the contribution of spectral and energy-based features to speakers' traits classification is assessed using the spectral part and an extension of the official feature set. Then the LOSO-cv results are given for the best joint contribution of spectral and prosodic feature sets. In section 5, the results on the Test set are presented. The last section concludes the study.

2. Baseline system

We developed a Baseline System (BS) from the description of the challenge baselines in INTERSPEECH 2012 Speaker Trait Challenge. For its development, we used the WEKA data mining tool kit for classification [14] of each seven traits of

the three sub-challenges. Following the baselines and for any trait T, each two-class classifier (T and NT) was based on -utterance representation by the official set of features (6,125 features), -Support Vector Machines (SVM) with linear Kernel. Synthetic Minority Over-sampling Technique (SMOTE) was used. The INTSINT targets were used to remove noise segments at the beginning and ending of the speech clips. The complexity parameter of the SVM classifier was optimized using Leave-One-Speaker-Out cross-validation (LOSO-cv) on the data (T&D) of both train (T) and development (D) sets. Taking into account speaker independency in the testing task, LOSO-cv on T&D was chosen as prediction criterion of classification performance. This process consists in splitting the data set into *s* speaker-dependant disjoint folds with *s* the number of speakers and to classify each instance of a fold from a model trained on the *s-1* other folds. The average performance obtained by such a computing process over all the folds warranty speaker independency of the observed results.

Table 1 gives the performances in Unweighted Accuracy rate (UA in percent) with the LOSO-cv using the 6,125 audio features of the Baseline System (BS) and the C^{BS} complexity parameter of the SVM classifier for each trait of the three sub-challenges tasks: O, C, E, A, N, L and I. The SVM complexity parameters (C^{BS}) were obtained in order to maximize the LOSO-cv on T&D. Two others LOSO-cv were computed: the first one on the Train set (T-cv), the second one on the Development set (D-cv). The following values of SVM complexity have been tested for the maximization of the LOSO-cv on T&D (T&D-cv): 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2 and 5. The classification results on the Development set from SVM trained on the Train set (T vs D) are also computed.

Table 1. *Baseline System results using the SVM complexity (C^{BS}) maximizing T&D-cv.*

Task	SVM C^{BS}	T vs D UA %	T-cv UA %	D-cv UA %	T&D-cv BS UA %
Personality Sub-Challenge – 6125 features					
O	0.02	60.5	73.8	73.2	**67.7**
C	0.01	73.2	74.3	72.5	**74.8**
E	0.02	80.4	72.0	79.8	**79.5**
A	0.001	64.9	54.8	58.5	**61.3**
N	0.005	67.6	72.3	63.4	**71.8**
Likability Sub-Challenge – 6125 features					
L	0.002	59.3	54.5	61.1	**59.8**
Pathology Sub-Challenge – 6125 features					
I	0.02	57.2	69.1	56.4	**63.9**

Two synergies are detected for C and A with the T&D-cv results higher than the T-cv and D-cv results. One inconsistency is detected for O with the T&D-cv result lower than the T-cv and D-cv results. It shows the difficulty to improve classification of this trait using more data.

3. Pitch and intonation for speaker trait

In the purpose of improving the extraction of prosodic information, we extended the official feature set. Then we used MOMEL (MOdelling Melody) and INTSINT (INternational Transcription System for INTonation) software [13] for modeling intonation contours. This Praat plugin was designed for analysis and generation of intonation patterns and used for synthesis of intonation in text-to-speech synthesis [15].

3.1. Pitch feature-based extension

The goal of the pitch feature-based extension is to obtain a better representation of the pitch curve. For the official acoustic feature set, TUM's open-source openSMILE [16] was used to compute the pitch (F0) using the SHS algorithm and Viterbi smoothing. Various functionals extracted 97 features (P1 set) from F0 and delta F0. An extended set of 134 features (P2 set) was defined using more functionals (e.g., statistics on amplitude peaks) described in Table 8. Another set of 134 features (P3 set) was also studied: P2 functionals were applied to the pitch computed without Viterbi smoothing.

3.2. MOMEL/INTSINT intonation modeling

MOMEL/INTSINT software [17] models intonation of an utterance by encoding the pitch curve into a sequence of labels aiming at the stylization of the curve into typical contours of intonation. MOMEL algorithm transcripts the smoothed pitch curve by quadratic spline function into a sequence of target points defined by couples (P_i, t_i) designing the value of pitch P_i at time t_i. From this sequence, a reduction procedure gives the targets detected as maximum of relevant local variation in the pitch curve and corresponding to major changes in the intonation contour.

Eight target labels are used for the contour stylization of the pitch curve: -three absolute labels from constant pitch value, T (Top), M (Medium), B (Bottom), and -five contextual labels from variable pitch values depending on the previous target, H (High: local maximum), U (Upstepped), S (Same as preceding), D (Downstepped), L (Low: local minimum). These labels are typical contours for the characterization of the intonation. In Table 2, the Low Level Descriptors (LLD) of the intonation model are described: the *key* and the *range* are two speaker/utterance-dependant descriptors followed by the values of the absolute and contextual target labels.

Table 2. *INTSINT low level descriptors for intonation model from [13].*

key	Mean value of pitch of the speaker's utterance
range	$[P_{min} .. P_{max}]$ of the utterance (in octave)
Absolute labels	
T	key*√(2^range)
M	key
B	key/√(2^range)
Contextual labels – P_{i-1}: pitch of the precedent target	
H	$\sqrt{(P_{i-1}*T)}$
U	$\sqrt{(P_{i-1}*\sqrt{(P_{i-1}*T)})}$
S	P_{i-1}
D	$\sqrt{(P_{i-1}*\sqrt{(P_{i-1}*T)})}$
L	$\sqrt{(P_{i-1}*B)}$

Four sets of functionals were developed to extract 34 features (I1) from the sequence of the target points defined by triplets (label, pitch value and time position). The first set was applied to target labels, the second to target pitch values, the third to the duration of the segments between two consecutive targets. The last set of functional was applied to the stylized pitch curve using ARMAX modeling [18]. Table 3

summarizes the four set of functionals applied to INTSINT low level descriptors.

Table 3. *Definition of functional applied to the INSINT low level descriptors.*

Functionals for INTSINT LLD
Functionals applied to target labels
percentage of the label in the sequence
Functionals applied to target pitch values
mean, variance, skewness, kurtosis
Functionals applied to segment between two targets
mean, variance of segment duration pitch rising: mean, variance of segment duration pitch falling: mean, variance of segment duration
Functionals applied to the stylized pitch curve
variance, linear regression slope, offset, normalized quadratic error (nqe) armax modeling coefficient (p = 1, q = 0) , a1, nqe armax modeling coefficients (p = 2, q = 0) , a1, a2, nqe armax modeling coefficients (p = 2, q = 1) , a1, a2, b1, nqe

3.3. Impact of pitch and intonation in performance

Table 4 gives the T&D-cv performances in UA of the pitch and intonation feature-based system using the optimal combination of features sets. For three traits – A, N, L – the UA rate is higher in absolute value of 2.3, 2.1 and 0.3 respectively than the baseline system. It is noticeable that 97, 134 and 168 pitch and intonation-based features respectively gave a better result than the 6,125 features of the official set. Pitch and intonation seem to be tied to specific traits.

Table 4. *Pitch and intonation feature-based system results using optimal combination and optimized SVM complexity (C^{PI}).*

Task	O	C	E	A	N	L	I
C^{PI}	5	2	2	5	1	5	5
Best set comb.	P3	P2+I1	P3	P1	P3	P3+I1	P3
# features	134	168	134	97	134	168	134
UA %	**62.2**	**71.9**	**72.9**	**63.6**	**73.3**	**60.1**	**58.9**
BS UA %	67.7	74.8	79.5	61.3	71.8	59.8	63.9

4. Feature selection for speaker trait

The goal is to obtain for each trait a specific feature set combining pitch and intonation related features with spectral features. In this purpose, two complementary features set were defined (S1 and S2) and their classification performance is assessed. Then we tested all the combinations of the feature sets (pitch, intonation, spectral) to obtain the best performance.

4.1. Spectral feature set

The first set (S1) was made up 6,028 features of the official set excluding the 97 features tied to pitch. The second set (S2) of 10,172 features was an extension of S1 using more functional described in Table 8. These features are mainly based on spectrum and energy LLD. Table 5 gives the T&D-cv performances in UA of the best spectral feature-based system. For five traits – O, C, A, L and I – the UA rate is higher in absolute value of 0.5, 0.3, 1.8, 0.6 and 1.4 respectively than the baseline system.

Table 5. *Best spectral feature-based system results using optimized SVM complexity (C^S).*

Task	O	C	E	A	N	L	I
C^S	0.02	0.01	0.01	0.001	0.005	0.005	0.01
Best set	S1	S1	S2	S2	S1	S2	S2
# features	6028	6028	10172	10172	6028	10172	10172
UA %	**68.2**	**75.1**	**79.5**	**63.1**	**71.5**	**60.4**	**65.3**
BS UA %	67.7	74.8	79.5	61.3	71.8	59.8	63.9

It is noticeable for three traits – A, L and I – this improvement is due to the extended spectral-based features. Table 6 gives the T&D-cv performances in UA of the pitch, intonation and spectral feature-based system using optimal combination of the features sets.

Table 6. *Best feature-based system results using optimized SVM complexity (C^B).*

Task	O	C	E	A	N	L	I
C^B	0.02	0.01	0.02	5	1	0.005	0.01
Best comb.	P3+S1	S1	P1+S1	P1	P3	P3+I1 +S2	S2
# features	6162	6028	6125	97	134	10342	10172
UA %	**68.4**	**75.1**	**79.5**	**63.6**	**73.3**	**61.1**	**65.3**
BS UA %	67.7	74.8	79.5	61.3	71.8	59.8	63.9

For six traits – O, C, A, N, L, I – the UA rate increased of 0.7, 0.3, 2.3, 1.5, 1.3, 1.4 respectively. It is noticeable that each trait has a specific feature set.

5. Test results

All systems are trained on both Training and Development sets. Table 7 presents the results achieved by the official baseline system (OBS) and our five submissions (Sub. #1 to #5) with their number of features.

Table 7. Submission results for Test sets classification.

Task	O	C	E	A	N	L	I
OBS UA%	59.0	79.1	75.3	64.2	64.0	59.0	68.9
# features Sub. #1	6162 57.0	6028 77.6	6125 75.8	97 56.5	134 59.9	10342 **63.7**	10172 **70.0**
# features Sub. #2	4782 55.4	3554 79.1	10057 76.2	10363 62.4	10209 **68.3**	8348 **64.1**	10006 69.9
# features Sub. #3	10297 55.3	10269 79.1	5664 75.1	10045 60.0	10073 **68.8**	10209 62.8	10362 69.3
# features Sub. #4	4281 56.0	467 77.6	7935 75.5	879 58.2	467 63.8	8397 50.0	9907 69.8
Fusion Sub. #5	57.5	80.1	74.6	61.0	**68.9**	**65.8**	69.5

Sub. #1 uses the features described in Table 6. The results showed an improvement of 4.7% on the Likability trait, an improvement of 1.1% on the Intelligibility trait and no improvement on the OCEAN traits compared to the official baseline. In Sub. #2, functionals and low level descriptors are selected by a Backward Best Fit algorithm (BFT) maximizing T vs D. The results showed an improvement on the OCEAN traits and on the Likability trait. In Sub. #3, BFT is used maximizing T&D-cv with similar results. In Sub. #4, all functionals are used and only low level descriptors are selected

by BFT. In Sub #5, fusion of submissions are computed on the OCEAN traits (from Sub. #2 and #3), on the Likability and Intelligibility traits (from Sub. #1, #2 and #3). The results showed an improvement on the Likability trait of 6.8% compared to the official baseline.

6. Conclusion

In this paper, we have presented trait classification system which was submitted to INTERSPEECH 2012 Speaker Trait Challenge. Six features sets were assessed: two sets were extracted from the official feature set and four sets were developed for this challenge. For each trait, a specific feature set was made up optimal combination of the six feature sets. These results have shown the pitch, intonation and spectral contribution to speakers' traits classification. Three kind of traits may be distinguished from the study, the traits tied to pitch – A and N – those tied to spectrum – C and I – and those tied to both pitch and spectrum – O, E, L. Intonation features were only useful for likability classification. Results on the Test set have shown a significant improvement of 6.8% on the likability trait, an improvement of 1.1% on the intelligibility trait and insignificant improvement on the OCEAN traits compared to the official baseline UA results.

Future works should include the development of new features related to INTSINT intonation model. Finally, the most frequent sequences of targets have to be taken into account using language modeling such as N-grams.

7. References

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

[2] Digman, J. M., "Personality structure: Emergence of the five-factor model", Annual Review of Psychology, Vol. 41, 417–440, 1990.

[3] Piedmont, R. L. and Aycock, W., "An historical analysis of the lexical emergence of the Big Five personality adjective descriptors", in Personality and Individual Differences, Vol. 42, N°6, 1059–1068, 2007.

[4] Pinto-Coelho, L., Braga, D., Sales-Dias, M. and Garcia-Mateo, C., "On the development of an automatic voice pleasantness classification and intensity estimation system", Computer Speech and Language, doi:10.1016/j.csl.2012.01.006, 2012.

[5] Gomez, A. M., Schwerin, B. and Paliwal, K., "Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio", Speech Communication, Elsevier, Vol. 54, 503–515, 2012.

[6] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. and Narayanan, S., "Paralinguistics in Speech and Language - State-of-the-Art and the Challenge", Computer Speech and Language (2010), doi:10.1016/j.csl.2012.02.005.

[7] Ketrow, S. M., "Attributes of a Telemarketer's Voice and Persuasiveness: A Review and Synthesis o/the Literature", in Journal of Direct Marketing, Vol. 4, N° 3, 7–21, 1990.

[8] Scherer, K. R., "Personality inference from voice quality: the loud voice of extroversion", European Journal of Social Psychology, Vol. 8, 467–487, 1978.

[9] Apple, W. and Krauss, R. M., "Effects of pitch and speech rate on personal attributions", Journal of Applied Social Psychology, Vol. 37, N° 5, 715–727, 1979.

[10] Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M. and Barry, W. J., "Modelling personality features by changing prosody in synthetic speech", in Proc. Speech Prosody, Dresden, Germany, 88–92, 2006.

[11] Rodero, E., "Intonation and Emotion: Influence of Pitch Levels and Contour Type on Creating Emotions", Journal of Voice, Vol. 25, No. 1, 25–34, 2010.

[12] Tanja Banziger, T and Scherer, K., R., "The role of intonation in emotional expressions", Speech Communication Vol. 46, 252–267, 2005.

[13] Hirst, D., "A Praat Plugin for MOMEL and INTSINT with Improved Algorithms for Modelling and Coding Intonation", in ICPhSXVI, 1233–1236, 2007.

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, vol. 11, 10–18, 2009.

[15] Véronis, J., Di Cristo, P., Courtois, F and Chaumette, C., "A stochastic model of intonation for text-to-speech synthesis", Speech Communication, Vol. 26, 233–244, 1998.

[16] Eyben, F., Wöllmer, M. and Schuller, B., "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", in Proc. ACM Multimedia. Florence, Italy: ACM, 1459–1462, 2010.

[17] Hirst, D., "Form and function in the representation of speech prosody". Speech Communication 46, n°3-4, 334–347, 2005.

[18] Hyndman, R.J. and Khandakar, Y., "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, Vol. 26, N°3, 76–78, 2008.

Table 8. *Additional functionals for F0 / Δ F0 and LLD / Δ LLD except F0*

Additional functionals for F0 / Δ F0
Functionals applied to F0 / Δ F0
max. , min., max. – mean, mean – min. geometric mean (gmean), duration
Functionals applied to F0 only
peak to peak ampl. diff.: mean, normalized mean, std.dev, normalized std.dev local min. peak: range, mean, mean–peak mean, mean/peak mean min. to min. ampl. diff.: mean, normalized mean, std.dev, normalized std.dev local min. to local min. rising slope : max., min. local max. to local max. falling slope : max., min.
Functionals applied to Δ F0 only
percentage of non-zero frames (nnz), nnz mean, nnz gmean absolute values : mean, nnz mean positive values : mean, quadratic mean root
Additional functionals for LLD / Δ LLD except F0
Functionals applied to LLD / Δ LLD
max. , min., mean, max. – mean, mean – min. quadratic mean percentage of non-zero frames (nnz) duration, number of detected segments
Functionals applied to LLD only
linear regression error, quadratic regression error peak to peak ampl. diff.: mean, normalized mean, std.dev, normalized std.dev local min. peak: range, mean, mean–peak mean, mean/peak mean min. to min. ampl. diff.: mean, normalized mean, std.dev, normalized std.dev local min. to local min. rising slope : max., min. local max. to local max. falling slope : max., min.
Functionals applied to Δ LLD only
values: nnz mean, nnz qmean, nnz geometric mean absolute values: mean, nnz mean positive values: quadratic mean root