INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Exploring Fusion Methods and Feature Space for the Classification of Paralinguistic Information

David Tavarez[1], *Xabier Sarasola*[1], *Agustin Alonso*[1], *Jon Sanchez*[1], *Luis Serrano*[1], *Eva Navas*[1], *Inma Hernáez*[1]

[1]AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain

{david,xsarasola,agustin,ion,lserrano,eva,inma}@aholab.ehu.eus

Abstract

This paper introduces the different systems developed by Aholab Signal Processing Laboratory for The INTERSPEECH 2017 Computational Paralinguistics Challenge, which includes three different subtasks: Addressee, Cold and Snoring classification. Several classification strategies and features related with the spectrum, prosody and phase have been tested separately and further combined by using different fusion techniques, such as early fusion by means of multi-feature vectors, late fusion of the standalone classifier scores and label fusion via weighted voting. The obtained results show that the applied fusion methods improve the performance of the standalone detectors and provide systems capable of outperforming the baseline systems in terms of UAR.

Index Terms: speech processing, classifier fusion, computational paralinguistics

1. Introduction

Computational Paralinguistics is a very wide field of research that every year gets the attention of a lot of researchers who take part in the ComParE challenges [1]. The 2017 challenge comprises three sub-challenges devoted to the detection of infant-directed speech, the identification of cold-speech and the classification of snoring sounds.

The infant-directed speech (IDS) has received a great attention during the last years because it is thought to have an important role in the child's development. For this reason, a lot of studies have tried to characterize IDS. Some studies are focused in the linguistic and prosodic characteristics of IDS [2], while others are centred in the phonetic differences between IDS and adult-directed speech [3] [4] [5]. In comparison, very few studies have addressed the acoustic characterization of cold-speech, mainly the work of Tull [6], who found significant differences in formant frequencies, nasality variables and cepstra between cold and healthy conditions. Finally, snoring is also a widely studied subject because of its relation with sleep disorders like apnoea [7]. As stated in [8], objective assessment of snoring is important to detect presence or absence of sleep-disordered breathing. This is the reason why there have been efforts like [9] and [10] to classify the snore sounds in an automatic way.

The advances made during the last years in feature extraction, classification and fusion techniques have allowed facing multiple problems applying similar approaches. The different problems posed in these three sub-challenges can be solved using several features related with the spectrum, prosody and phase of the signals and applying diverse classification strategies. This is precisely the approach followed in this work.

2. Magnitude spectrum features

The features related to the module of the short-time spectrum are probably the most used information for speech signal processing. We have employed two different features derived from this magnitude: canonical MFCCs and the recently proposed CQCCs. Both of them reflect the overall structure of the power spectrum envelope in short quasi-stationary analysis frames, but they enhance different frequencies of the Fourier spectrum.

2.1. MFCC

The MFCC magnitude-based features [11] are calculated using 12 Mel-frequency cepstral coefficients without the zeroth energy coefficient. Their first and second derivative values are also computed leading to 36 parameters calculated every 10 ms.

2.2. CQCC

The constant Q cepstral coefficients (CQCC) [12] are based on the coupling of the constant Q transform (CQT) with traditional cepstral analysis. The constant Q factor across the entire spectrum results in a higher frequency resolution at lower frequencies while providing a higher temporal resolution at higher frequencies, a close reflection of the human perception system. For these experiments, 90 parameter vectors were selected.

3. Relative phase shift features

The RPS is a representation for the harmonic phase information described in [13]. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency. The RPS representation consists in calculating the phase shift between every harmonic and the fundamental component at a specific point of the fundamental period, namely the point where the instantaneous phase of the fundamental component (φ_o) is 0.

$$\Psi_k(t_a) = \varphi_k(t_o) = \varphi_k(t_a) - k\varphi_1(t_a) \quad (1)$$

Equation (1) defines the RPS transformation which allows computing the RPS (Ψ_k) of the k-th harmonic from the instantaneous phases (φ_o, φ_k) at any point (t_a) of the signal. The RPS values are wrapped to the $[-\pi, \pi]$ interval. The RPS values are not well suited for statistical modelling, so to create and test the models the socalled DCT-mel-RPS parametrisation is used instead. To obtain the parameters, the differences of the unwrapped RPS values are filtered with a mel filter bank (48 filters) and a discrete cosine transform (DCT) is applied to the resulting sequence. The DCT is truncated to 20 values and the first and second derivative values are calculated.

For the experiments, the speech signals are windowed every 10 ms (using hamming windows of a length of 3 pitch periods)

http://dx.doi.org/10.21437/Interspeech.2017-1378

and the RPSs are calculated from the Fourier spectrum only for voiced frames. Then the DCT-mel-RPS parametrisation is applied to every frame and the averaged value of the slope of the unwrapped RPS values is also included which leads to a total of 63 phase-based parameters.

4. Suprasegmental features

Suprasegmental features represent long-term information, estimated over time intervals longer than a frame, generally the time between two consecutive pauses. In this work we used a whole utterance as integration time.

4.1. Spectral statistics

Log-Filter Power Coefficients are known to outperform traditional MFCC or LPCC parameters in cases like emotion identification [14], so they were chosen for the frame-wise spectral characterization. LFPC features represent the spectral envelope in terms of the energy in Mel-scaled frequency bands.

Long-term log-filter power coefficients (LLFPC) were calculated for the suprasegmental characterization of the spectrum. For each of the 18 LLFPC coefficients and their first and second derivatives six statistics were computed: mean, variance, minimum, range, skewness and kurtosis. At the end, 18x3x6=324 suprasegmental spectral features were extracted.

4.2. Voice quality features

Five features related to voice quality (LVQ) were computed only for vocalic segments, in order to consider only segments with reliable glottal source estimation: jitter and shimmer estimated using the five-point period perturbation quotient (ppq5) and five-point amplitude perturbation quotient (apq5) values as defined in Praat, mean normalised amplitude quotient calculated by averaging the normalised amplitude quotient obtained all along the vowel and mean spectral tilt and mean spectral balance all along the vowel. The values corresponding to vowels in the same integration segment were then averaged in order to obtain a single feature vector for the whole integration time.

The vocalic segments were automatically detected using the HMM based phoneme recogniser described in [16].

4.3. Prosody

The considered prosodic features (LPROS) are divided into five categories, according to the nature of the information they represent. Altogether 54 prosodic features have been defined.

- Intonation Statistics: F0 value was estimated every 10 milliseconds with the algorithm described in [15] and first and second derivatives were calculated. For each of these three curves, the same six statistics used for the spectral characterization were computed to produce 18 intonation-related features.
- Intensity Statistics: First and second derivatives were calculated for the power curve and the six statistics were computed, giving another 18 power-related features.
- Speech Rate: It has been characterised as the mean and variance of vowel duration.
- Regression Features: In each detected vowel, a linear regression was estimated for the values of F_0 and intensity. Then the absolute value of the regression line slope was calculated, and six additional features were extracted: mean, variance and maximum of absolute values of intonation and power slopes in vowels.
- Sentence-End Features: Five more features were extracted from the last vowel detected in the integration time: slope of its intonation and intensity, central value of its intonation and intensity and duration. Normalised and unnormalised versions of these five values were considered. The normalised values are defined as the corresponding non-normalised ones divided by the mean value over all the vowels detected in the integration segment.

5. Music related features

Interest in music information retrieval has been growing in late years because of the need of digital music services. This research area is creating new methods to classify music and new features are been searched and created to help in this task. Although these are features prepared for music signals, the distribution of power in the spectrum, harmonics and tone analysis are parameters that can be applied to any sound signal. In this work we used chroma, spectral contrast and tonal centroid features extracted with the python library Librosa [17].

5.1. Chroma vector

Chroma features (CHR), more commonly known as Pitch Class Profiles (PCP), are the representation of the energy of a signal in predefined pitch classes (usually 12 classes due to the western tonality system). It was first introduced in [18] to analyse music because of its robustness to changes in timbre and instrumentation.

5.2. Tonal centroid

Tonal centroid features (TON) are six-dimensional feature vectors used to analyse harmonic changes in music [19]. They are based on the Harmonic Network or Tonnetz, a representation of the tonal space in a lattice diagram with fifths and major/minor third relations that helps to find tonal distance and relationships.

5.3. Spectral contrast

Spectral contrast (SC) is an extension of the MFCC algorithm to analyse music signals [20]. It takes into account the spectral peak, spectral valley and their difference in each sub-band. SC feature could roughly reflect the relative distribution of the harmonic and non-harmonic components in the spectrum.

6. Classifiers

Two different modelling approaches have been used depending on whether the feature extraction involves short or long term statistics. Short term spectral and music related features have been used to train different Gaussian mixture models (GMM) for each class in the database. On the other hand, a cosine distant classifier has been applied to the long term features extracted from a whole utterance. Only the training dataset has been used to train these classifiers.

6.1. Short term statistics: GMM

A GMM based binary classifier has been used to retain the information of the audio signals according to the different subtasks. First, during the training phase, GMM parametric models are created for all the classes in the database (C-NC, ADS-IDS...).

To perform the posterior classification task, the system tests a candidate vector sequence against the different speech models to get the corresponding likelihood values. The candidate is considered to belong to one of the classes if the likelihood ratio between the target classes exceeds a certain decision threshold, which was set to maximise the UAR value over the development set. During development tests, different mixture models were evaluated to select the appropriate mixture number for the system and the use of 1024 Gaussian mixtures was finally set for the experiments.

6.2. Long term statistics: cosine distance

In the long term parametrisation a single feature vector is extracted from each recording. To classify these vectors a cosine distance classifier has been used in this work. The cosine distance of a target and a test vector is defined as the normalised dot product of both vectors. For every audio file in the database a single vector is extracted and its cosine distances to all target vectors representing the different classes are calculated. Again, the candidate is classified based on a certain threshold which was set during the tuning process. The target vectors representing the classes have been calculated by averaging all the vectors of the training audio for a certain class. Previously all the vectors have been whitened by subtracting the global mean of the training material, scaled by the inverse square root of the global covariance matrix and normalised to the unit length.

6.3. i-vector transformation

In order to use the short term information along with the cosine distance classifier, i-vectors have been extracted from the spectral, phase and music related features. To compute the i-vectors from each type of features, first a UBM has been trained. We used 10 EM iterations to train a G_{mix} F-dimensional GMM. For each case, the dimension F corresponds to the size of the feature vector extracted in the parametrization step. Also G_{mix} was optimised to maximise the results obtained in the development set for each different parametrisation. This GMM was then used to compute the zero and first-order statistics for all the training instances. The total variability matrix T was randomly initialised and estimated from the sufficient statistics with 10 iterations of EM. Once T matrix was trained, we extracted M-dimensional i-vectors for all the instances in the database. This procedure was repeated for different values of i-vector dimension and empirically set to 50 based on the experimental results on the development part of the database.

7. Fusion

The information provided by the different classifiers has been combined to produce a final decision using two different fusion strategies, namely early fusion and late fusion. In early fusion, for every utterance the vectors extracted from each audio file are combined into a single multifeature vector. This vector is used in a cosine distance based classifier which produces a new and more accurate score. In late fusion, standalone classifiers are trained with the different vectors calculated from all the features. Their scores are then combined using a weighted sum to calculate the final score. In this work, the weights have been learned over the development subset of the training material using the Bosaris toolkit [21].

Table 1: *Results obtained for the Cold detection in terms of UAR and Accuracy for the development set*

Feature	Acc.	UAR
Standalone GMM		
MFCC	53.9	64.7
CQCC	59.2	**65.7**
RPS	75.2	54.4
CHR	68.6	60.6
TON	56.9	61.0
SC	77.5	62.2
Standalone cosine distance		
MFCC	84.2	**63.4**
CQCC	82.5	60.6
RPS	62.4	60.5
LLFPC	74.1	60.7
LVQ	75.8	54.1
LPROS	68.5	58.7
TON	67.4	58.4
SC	64.2	61.5
Early vector fusion		
MFCC+CQCC	82.2	64.1
MFCC+CQCC+RPS	83.5	64.8
MFCC+CQCC+TON+SC	84.2	65.1
MFCC+CQCC+RPS+TON+SC	82.3	**65.1**
Late fusion cosine distance		
MFCC+CQCC+RPS	82.8	64.1
MFCC+CQCC+RPS+TON	83.0	64.5
MFCC+CQCC+RPS+TON+SC	75.5	**64.7**
Late fusion GMM		
MFCC+CQCC	51.1	65.8
MFCC+CQCC+TON	55.0	66.9
MFCC+CQCC+TON+CHR	52.4	**67.0**
Late fusion GMM & cosine distance		
$CS_{MFCC+CQCC+RPS+TON+SC}$+$GMM_{MFCC}$	56.2	66.1
$CS_{MFCC+CQCC+RPS+TON+SC}$+$GMM_{MFCC+CQCC}$	64.5	66.7
$CS_{MFCC+CQCC+RPS+TON+SC}$+$GMM_{MFCC+CQCC+TON}$	61.7	**69.1**
Baseline systems		
Best standalone system	-	64.0
System with best development results	-	**66.1**
System with best test results	-	65.2

8. Results

Table 1 shows the results obtained for the Cold detection task. The results for the standalone systems are listed in the first place. MFCC and CQCC parameters show better results in both modelling approaches, with CQCC outperforming all the other features when working independently. Voice quality and prosody are the features with the worst behaviour for this task, even in the fusion step, with no improvement achieved in any case. Both early and late fusion improve the performance of the standalone systems and obtain good results, above the provided baselines systems. This good behaviour remains in the test set,

Table 2: *Results obtained for the Addressee task in terms of UAR and Accuracy for the development set*

Feature	Acc.	UAR
Standalone GMM		
MFCC	58.8	**63.0**
CQCC	58.1	59.0
RPS	58.0	59.9
CHR	68.6	60.6
TON	58.5	59.7
SC	56.8	59.0
Standalone cosine distance		
MFCC	57.4	60.1
CQCC	54.2	57.8
RPS	57.8	58.2
LLFPC	60.6	57.6
LVQ	58.1	57.2
LPROS	58.5	**60.2**
TON	58.9	55.8
SC	56.6	58.8
Early vector fusion		
RPS+LPROS	58.5	60.6
MFCC+LPROS	58.5	**61.1**
SC+LPROS	62.0	60.3
Late fusion cosine distance		
MFCC+RPS	60.8	61.1
MFCC+RPS+TON	61.6	61.9
MFCC+CQCC+RPS+TON+LPROS.	62.3	**63.7**
Late fusion GMM		
MFCC+CQCC	59.2	63.2
MFCC+CQCC+RPS	63.0	64.0
MFCC+CQCC+RPS+TON	63.9	64.4
MFCC+CQCC+RPS+TON+SC	61.9	64.6
MFCC+CQCC+RPS+TON+SC+CHR	63.9	**65.0**
Late fusion GMM & cosine distance		
CS_{MFCC}+$GMM_{MFCC+CQCC+RPS+TON+SC+CHR}$	64.4	65.2
$CS_{MFCC+LPROS}$+$GMM_{MFCC+CQCC+RPS+TON+SC+CHR}$	65.1	**66.0**
Baseline systems		
Best standalone system	-	61.8
System with best development results	-	**67.8**
System with best test results	-	66.4

Table 3: *Results obtained for the Snoring classification task in terms of UAR and Accuracy for the development set*

Feature	Acc.	UAR
Standalone cosine distance		
MFCC	37.8	46.9
CQCC	33.9	39.4
RPS	33.6	30.5
CHR	35.0	35.9
TON	25.8	38.0
SC	41.7	**47.2**
Early vector fusion		
MFCC+SC	37.5	**47.5**
MFCC+SC+CQCC	33.2	43.2
Late fusion cosine distance		
MFCC+RPS	51.2	**54.3**
Baseline systems		
Best standalone system	-	40.6
System with best development results	-	**46.6**
System with best test results	-	40.6

where a 66.2 UAR value has been obtained.

Table 2 shows the results obtained for the Addressee task. In this case prosodic features show better results than in the previous task, although MFCC outperforms all the other features when working independently. Voice quality and prosodic parameters improve the results obtained by the spectral features when the different fusion techniques are applied. Both early and late fusion improves the performance of the standalone systems and obtain good results, with a 65.5 UAR value in the test set.

Finally, Table 3 shows the results obtained for the snore classification task. Only MFCC and contrast features perform an accurate snore classification. In this case also early and late fusion improve the performance of the standalone systems. Despite the poor classification provided by some of the proposed features, a 50.6 UAR value has been obtained in the test set.

Results for the test set in all the tasks do not reach those of the baseline, due to the simplicity of the classifiers and the limitations of the fusion process. However, the features applied contain valid information to address these tasks.

9. Conclusions

In this work we have introduced the different systems developed by Aholab Signal Processing Laboratory for The INTERSPEECH 2017 Computational Paralinguistics Challenge, which includes three different subtasks: Addressee, Cold and Snoring classification. Several classification strategies and features have been evaluated and different fusion techniques have been tested.

The obtained results show that spectral parameters perform better in the cold detection task, while prosodic features are capable of improving the results of the standalone classifiers in the addressee one. Only MFCC and contrast features show certain capabilities to classify the snore signals correctly.

The applied fusion methods widely improve the performance of the standalone detectors and provide systems capable of outperforming the baseline systems in terms of UAR in some cases. Furthermore, the fusion of these systems with the baseline system proposed by the organisers is expected to improve the results.

10. Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER support (RESTORE project,TEC2015-67163-C2-1-R) and the Basque Government (ELKAROLA project, KK2016/00087).

11. References

[1] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "Paralinguistics in speech and languageState-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] C. Saint-Georges, M. Chetouani, R. Cassel, F. Apicella, A. Mahdhaoui, F. Muratori, M.-C. Laznik, and D. Cohen, "Motherese in interaction: at the cross-road of emotion and cognition?(a systematic review)," *PloS one*, vol. 8, no. 10, pp. 1–17, 2013.

[3] A. Cristià, "Phonetic enhancement of sibilants in infant-directed speech," *The Journal of the Acoustical Society of America*, vol. 128, no. 1, pp. 424–434, 2010.

[4] F. Pons, J. C. Biesanz, S. Kajikawa, L. Fais, C. R. Narayan, S. Amano, and J. F. Werker, "Phonetic category cues in adult-directed speech: Evidence from three languages with distinct vowel characteristics." *Psicologica: International Journal of Methodology and Experimental Psychology*, vol. 33, no. 2, pp. 175–207, 2012.

[5] L. C. Dilley, A. L. Millett, J. D. McAuley, and T. R. Bergeson, "Phonetic variation in consonants in infant-directed and adult-directed speech: the case of regressive place assimilation in word-final alveolar stops," *Journal of Child Language*, vol. 41, no. 01, pp. 155–175, 2014.

[6] R. G. Tull, *Acoustic analysis of cold-speech: Implications for speaker recognition technology and the common cold*. Northwestern University, 1999.

[7] A. M. Alencar, D. G. V. da Silva, C. B. Oliveira, A. P. Vieira, H. T. Moriya, and G. Lorenzi-Filho, "Dynamics of snoring sounds and its connection with obstructive sleep apnea," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 1, pp. 271–277, 2013.

[8] D. Pevernagie, R. M. Aarts, and M. De Meyer, "The acoustics of snoring," *Sleep medicine reviews*, vol. 14, no. 2, pp. 131–144, 2010.

[9] A. Yadollahi and Z. Moussavi, "Automatic breath and snore sounds classification from tracheal and ambient sounds recordings," *Medical engineering & physics*, vol. 32, no. 9, pp. 985–990, 2010.

[10] C. Doukas, T. Petsatodis, C. Boukis, and I. Maglogiannis, "Automated sleep breath disorders detection utilizing patient sound analysis," *Biomedical Signal Processing and Control*, vol. 7, no. 3, pp. 256–264, 2012.

[11] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '83.*, vol. 8, 1983, pp. 93–96.

[12] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey 2016, The Speaker and Language Recognition Workshop*, 2016.

[13] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, March 2009.

[14] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603 – 623, 2003.

[15] I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–1057–IV–1060.

[16] I. Luengo, E. Navas, J. Sanchez, and I. Hernaez, "Detección de vocales mediante modelado de clusters de fonemas," *Procesamiento del Lenguaje Natural*, vol. 43, no. 0, pp. 121–128, 2009.

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015.

[18] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music." in *ICMC*, 1999, pp. 464–467.

[19] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, 2006, pp. 21–26.

[20] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1, 2002, pp. 113–116.

[21] E. de Villiers and N. Brummer, "Bosaris toolkit." [Online]. Available: https://sites.google.com/site/bosaristoolkit/