Design and evaluation of reduced-size feature sets for the assessment of sincerity in speech

E. M. Albornoz[1,2*] and C. E. Martínez[1,3]
[1]Research Institute for Signals, Systems and Computational Intelligence (sinc(i))
Dept. of Informatics, Faculty of Engineering and Water Sciences, University of Litoral
CC217, Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina
[2]National Scientific and Technical Research Council (CONICET), Argentina
[3]Cibernetics Lab, Faculty of Engineering, University of Entre Ríos
* emalbornoz@sinc.unl.edu.ar

***Abstract*—The recognition of states and traits of speakers is a significant issue to investigate, to be able to achieve more useful interactive systems. The sincerity of a speaker is a relevant paralinguistic phenomenon, which have not received too much attention from the affective computing community. In this work, we tackle the problem using novel feature sets proposed for emotion recognition. In addition, bioinspired features (using an auditory signal representation) and other spectral features are also evaluated. Finally, diverse combinations of these reduced-size feature sets are built. The provided standard, complete set with 6373 features is used for comparison purposes. Results show that using the combination of the proposed representations and state-of-art features, it is possible to obtain very small feature sets (less than 3% of the original size) that get comparable correlation measure with respect to the baseline.**

***Index Terms*—sincerity recognition, auditory representation, reduced-size feature sets**

and to

I. Introduction

In the last years, the affective computing community is dealing with new challenges looking to improve the emotional human-computer interaction. Today, one of the most relevant modalities is the recognition of emotions in speech. Recently, a new task was proposed in this context, namely the recognition of the sincerity of the speaker [1]. To address this challenge, literature shows that several vocal cues can be extracted to detect closely related speech acts like the sarcasm and verbal irony. Then, the tempo, intensity, pitch, local and global prosodic information, spectral features, among others are usually measured [2–4]. Some evidence about the differences between sarcastic and sincere intonation can be mainly found in the speaking rate, intensity and general hyperarticulation, which also can vary between languages, for example English and French [5,6]. These features have been used to build robust spoken dialogue systems, which are able to learn and detect the presence of sarcasm [7,8].

In this work, we build and evaluate the feasibility of tackling the task using very small feature sets. We compute two state-of-art minimalistic feature sets (GeMAPS and eGeMAPS) [9] and a set of spectral characteristics proposed in [10]. In addition, we propose a set of features based on a bioinspired model, computed by the auditory model proposed by [11]. As this model tries to mimic the auditory system, it is interesting to know if the model properties are useful for the recognition of sincerity. It is important to note that this model has been useful in feature extraction for related tasks in robust speech and emotion recognition [12–14]. Furthermore, some combinations of these feature sets are evaluated.

The organization of this paper is as follows. In Section II, brief descriptions of the speech corpora and baseline system are given. Next, state-of-art features and our methods for feature extraction are described. Section III presents the results obtained along with a discussion about the usability of the proposed features. Finally, Section IV gives the general conclusions and outlines future work.

II. Materials and methods

This section resumes the speech database, the baseline systems on the task and our approach to feature extraction.

A. Speech data and baseline system

The dataset provided is the *Sincerity Speech Corpus* (SSC) provided by the Columbia University and consists of two sets (train and test partitions) containing the utterances of 22 subjects (655 instances) and 10 subjects (256 instances), respectively. The recordings correspond to people reading six sentences expressing apologies in four different prosodic styles. The sentences vary in length, from one word ("Sorry") up to long phrases (for example "I am sorry, but I am going to have to decline your generous offer. Thank you for considering me."). The complete set has approximately 72 minutes of speech. Each instance was rated by a group of 13-19 annotators in terms of perceived sincerity using a scale from 0 (not sincere at all) to 4 (extreme sincere). For more details, we refer the reader to [1].

A state-of-art feature set is obtained from the speech recordings using the *openSMILE* toolkit [15], the so-called ComParE feature set. It calculates 6373 acoustic features using diverse functionals over low-level descriptor (LLD) contours. A full description can be found in [16].

The baseline system provided for the task consists of Support Vector Regression (SVR) with linear kernels and epsilon-insensitive loss. The system was trained with the Sequential Minimal Optimization (SMO) algorithm available in the open-source machine learning software WEKA. More details can be found in [1].

B. GeMAPS and extended GeMAPS

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) consists of a basic standard set intended to be used in various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis. In an interdisciplinary work, the creators presented the set as a minimalistic group of voice parameters (contrary to others large sets), with a selection based on their evidence to index affective changes in voice production, the feasibility to build algorithms for automatic calculations and their theoretical significance. One main objective was to bring to the affective computing community a common baseline for future evaluation of systems, so eliminating the differences due to internal parameters or even implementations of the same features across the research groups.

The GeMAPS consists of 62 parameters originated from 18 LLD descriptors divided into the following categories:

- frequency: pitch and its jitter, and first 3 formants;
- energy/amplitude: shimmer, loudness and harmonics-to-noise ratio;
- spectral: different ratios and indices showing relations between energy bands and peaks.

The extended GeMAPS (eGeMAPS) is an alternative version which adds 26 extra parameters to the basic set. They are obtained from cepstral coefficients along with dynamic information. The implementation of GeMAPS is publicly available with the openSMILE toolkit. Full details of the sets can be found in [9].

C. Mean of log-spectrum

In previous works we proposed the Mean of Log-Spectrum (MLS) coefficients, a set of features calculated from spectral data for different frequency bands. They were thought as an extra process to extract prosody information and were first used in the analysis and characterization of spoken emotions, in clean and noisy conditions [10,17]. Briefly, the MLS coefficients are defined using the signal spectrogram

$$S(k) = \frac{1}{N}\sum_{n=1}^{N} \log|v(n,k)|, \qquad (1)$$

where k is a frequency band, N is the number of frames in the utterance and $v(n,k)$ is the discrete Fourier transform of the signal in the frame n. For the computation, the spectrograms were obtained with Hamming windows of 25 ms. The first 30 MLS coefficients, corresponding to lower frequencies ($0-1200$ Hz), were considered based on evidence that the most useful information for emotion recognition was found in this frequency interval.

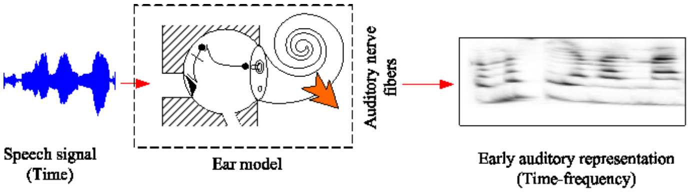

Fig. 1. Conceptual scheme for obtaining the auditory spectrogram

This set of 30 features is also considered along with the standard deviation of each coefficient, giving a second alternative set of 60 features.

D. Mean of the log-auditory spectrum

The representation of the sound signal at the cochlear level and auditory cortical areas has been studied as an alternative to classical analysis methods, given its intrinsic selective tuning to relevant natural sound [18]. Here, additionally to the described MLS coefficients, we propose to analyze the speech utterances by means of a related set of features based on the auditory spectrogram.

In [19], a model based on neurophysiological investigations at various stages of the auditory system was proposed. This model consists of two consecutive stages: first, it obtains an early auditory spectrogram with the activity of auditory nerve fibres, and then a model of the primary auditory cortex is used to process the spectrogram and find the spectro-temporal receptive fields. The model first process the speech signal with a bank of 128 cochlear (bandpass) filters in the range $[0-4000]$ Hz. The centre frequency of the filter at location x on the logarithmic frequency axis (in octaves) is defined as

$$f_x = f_0 2^x (\text{Hz}) \qquad (2)$$

where f_0 is a reference frequency of 1 kHz. This frequency distribution proved to be satisfactory for the discrimination of acoustic clues in speech and further reconstruction of the signals [20]. One important aspect is the fact that the first 71 coefficients correspond to the $[0-1220]$ Hz interval. Given the previous evidence for the MLS coefficients that the most useful information for emotion recognition is precisely found in this frequency interval [10], only this range is considered here.

After filtering, the second stage transduces the outputs into auditory-nerve patterns using a high-pass filter (modeling the fluid-cilia coupling), a non-linear activation function (compression in the ionic channels), and a low-pass filter (hair-cell membrane leakage). Then, the lateral inhibitory network is modeled by the half-wave rectification of the first-order derivative of the frequency. The output on each frequency band is finally obtained by integrating this signal over a short window [21]. Figure II-D shows a block diagram model that conceptually summarizes this procedure.

Considering the output of the first stage of the model, an alternative set of features is built using the mean of the log auditory spectrogram (MLSa), as

$$S_a(k) = \frac{1}{N}\sum_{n=1}^{N} \log |a(n,k)|, \tag{3}$$

where k is a frequency band, N is the number of frames in the utterance and $a(n,k)$ is the k-th coefficient obtained by applying the auditory filter bank to the signal in the frame n. The MLSa were computed using auditory spectrograms calculated for windows of 25 ms without overlapping.

The number of MLSa features is 71, which is also doubled similar to MLS by adding the standard deviation of each coefficient (142 features).

E. Support vector machines for function estimation

Support Vector Machines (SVM) are well-stablished learning models focused on minimizing the structural risk based on the available training data, in order to get a good generalization of these patterns. In the case of the sincerity task, the goal is to evaluate the fitting of the test results to a real scale given by the perceived sincerity (ground truth). The Support Vector algorithm is then used to carry out Support Vector Regression (SVR) by the estimation of a function $f(x)$ that minimizes the empirical risk

$$R_{\text{emp}}[f] = \frac{1}{l}\sum_{i=1}^{l} c\big(x_i, y_i, f(x_i)\big)\big) \tag{4}$$

for the training data $\boldsymbol{X} = \big[(x_1, y_1), \ldots, (x_l, y_l)\big]$ and the cost function $c(x, y, f(x))$. When dealing with overfitting, the minimization given is addressed as a regularization problem [22].

For the sake of comparison between our proposed feature sets and the baseline, we use the same SVM for regression that the baseline, that is, a SV regression where the cost function considered is the ϵ-insensitive loss $c(x, y, f(x)) = |y - f(x)|_\epsilon$ (robust against overfitting). Similar to the baseline, the training is carried out using the SMO algorithm for solving quadratic programming problems [23].

III. Results and discussions

This section gives the details of the experiments carried out. We first present the tuning and evaluation of the system on a leave-one-speaker-out cross-validation (LOSO-CV) scheme on the training data. Once the best configurations are found, the system is trained using all training data and thus the test set performance is obtained.

The numerical experiments were carried out using the SVR with linear kernels and fixed $\epsilon = 1.0$ with the WEKA software. The complexity parameter C of the SVR was optimised in this stage (for brevity reasons, only the best result in each case is reported). After the tuning phase, the results on the test set are presented and discussed.

The Spearman's Correlation Coefficient ρ was used as the figure of merit. It quantifies the extend of statistical dependence between a pair of observations, and constitutes a robust alternative to Pearson's correlation coefficient.

TABLE I
Spearman correlation coefficient (ρ) obtained for different sets.

Feature set	C	# of features	ρ
Baseline	10^{-4}	6373	0.4743
GeMAPS	10^{-1}	62	0.4128
eGeMAPS	10^{-1}	88	0.4504
MLS (mean)	10^{-1}	30	0.2254
MLS (mean+std)	1.0	60	0.3203
MLSa (mean)	10^{-1}	71	0.3124
MLSa (mean+std)	1.0	142	0.2712
MLS+MLSa (mean)	10^{-1}	101	0.2708
MLS+MLSa (mean+std)	10^{-1}	202	0.2767

Table I shows the obtained results (ρ coefficient) on the LOSO-CV scheme using the different feature sets. For each set, it can be seen the number of features and the complexity. The two last rows represent the aggregation of MLS and MLSa feature sets. As can be observed, the baseline (complete) feature set reaches the best performance. However, the eGeMAPS set has a well suited behaviour using just 88 values (1.4% of the number of features in the baseline set). On the other hand, in this first series of experiments, all our proposed sets get a lesser performance than baseline and GeMAPS.

A second group of experiments was done in order to evaluate the behaviour of MLS and MLSa sets aggregated with baseline, GeMAPS and eGeMAPS features. Although we pursue the best performance using low dimension sets, the combination of MLS and MLSa with baseline set is useful for comparison purposes. Table II shows the best results obtained on the training set using the same LOSO-CV scheme. The rows are divided into three groups depending on which are the combined sets. The second column indicates the complexity of the best SVR model while the third column shows the number of features in the set. Here, the results that improve the baseline are evidenced with a bold face. The absolute best result is also grayed, and it was reached using eGeMAPS+MLSa features.

As can be seen, the different combinations of baseline features with MLS/MLSa obtain just a minor improvement, in the best case from 0.4743 to 0.4750 (baseline+mean values of MLS and MLSa together). Instead, we can see that the combinations of GeMAPS/eGeMAPS with our proposal of features perform much better: for GeMAPS+mean MLSa the ρ coefficient rises from the baseline 0.4743 up to 0.4987, while in the case of eGeMAPS+mean MLSa it reaches 0.5232. On the other hand, the standard deviations (for both MLS and MLSa) are not providing any influential information given that, in almost all the cases, the obtained correlation declines. After this first analysis, we chose three models with different features and complexity, for the evaluation on the test set. The selected systems are marked with "$\rightarrow s$" in Table II.

Table III presents the ρ coefficients obtained for the selected models using the training and test data. It can be observed that the higher performance on the test data is reached using the complete feature set (baseline), which obtains ρ=0.602. Our best result using the test data is the

TABLE II
SPEARMAN CORRELATION COEFFICIENT (ρ) FOR COMBINED FEATURE SETS.

Feature set	C	# of features	ρ
Baseline	10^{-4}	6373	0.4743
Baseline features with			
MLS (mean)	10^{-4}	6373+30	**0.4744**
MLS (mean+std)	10^{-4}	6373+60	0.4737
MLSa (mean)	10^{-4}	6373+71	**0.4746**
MLSa (mean+std)	10^{-4}	6373+142	0.4718
MLS+MLSa (mean)	10^{-4}	6373+101	**0.4750**
MLS+MLSa (mean+std)	10^{-4}	6373+202	0.4716
GeMAPS features with			
MLS (mean)	1.0	30+62	0.4459
MLS (mean+std)	1.0	60+62	**0.4754**
MLSa (mean)	1.0	71+62	**0.4987**
MLSa (mean+std)	1.0	142+62	**0.4744**
MLS+MLSa (mean)	1.0	101+62	**0.4878**
MLS+MLSa (mean+std)	10^{-1}	202+62	0.4420
eGeMAPS features with			
MLS (mean) $\rightarrow s$	10^{-1}	30+88	0.4513
MLS (mean+std)	10^{-1}	60+88	0.4362
MLSa (mean) $\rightarrow s$	1.0	71+88	**0.5232**
MLSa (mean+std)	10^{-1}	142+88	0.4578
MLS+MLSa (mean) $\rightarrow s$	10^{-1}	101+88	**0.5093**
MLS+MLSa (mean+std)	10^{-1}	202+88	0.4252

TABLE III
SPEARMAN CORRELATION COEFFICIENT (ρ) FOR THE BEST COMBINATIONS OF SETS (USING THE PREVIOUSLY REPORTED C VALUES).

Feature set	# of features	ρ	
		Train	Test
Baseline	6373	0.4743	**0.602**
eGeMAPS+MLS	30+88	0.4513	0.5278
eGeMAPS+MLSa	71+88	**0.5232**	0.4594
eGeMAPS+MLS+MLSa	101+88	0.5093	0.4923

combination of eGeMAPS+MLS, which reaches a good correlation of $\rho = 0.5278$ (underlined number). Also, it can be seen that the best combination of features previously found, the eGeMAPS+MLSa, slightly reduces the correlation from 0.5232 to 0.4594 on the test data. The remaining combination of the three types of features considered, eGeMAPS+MLS+MLSa, nearly maintains the correlation.

Despite the fact that the feature sets proposed in this work obtained lower performances on the test data, some interesting points can be observed on training data. First, in the comparison GeMAPS vs. eGeMAPS the cepstral and dynamic information provided in the eGeMAPS is beneficial for the considered task. Second, the results obtained using the auditory analysis make evident its asserted natural ability to capture important, discriminative information contained in the speech signal. Here, the auditory spectral and prosody information (MLSa features) is providing useful information with respect to both: the analysis given by the MLS features and the full set of measurements given by the baseline features. This conclusion is also supported by the improvement of ρ when combining eGeMAPS with MLS or MSLa features: the correlation grows from 0.4513 to 0.5232 (see Table II).

On the other side, the sizes of the feature sets are reduced to a large degree. There is a considerable improvement in the performance on the training data ($\rho = 0.4743$ to 0.5232) using only 159 features given by 88 eGeMAPS + 71 MLSa coefficients, which represents about 2.5% of the original number of features. On the test data, the best performance is obtained with 88 eGeMAPS + 30 MLS features. Here, this set of 188 coefficients represents a great improvement in the size reduction to just 1.88% of the original size. The ability of the features to capture relevant information and the uniqueness of each one of the selected parameters in the representation, could be the factors that lead to obtain comparable performances facing the unaddressed challenge of evaluating the sincerity in a speech utterance.

IV. CONCLUSIONS

In this work, we proposed new reduced-size feature sets for the evaluation of sincerity in speech, which is a paralinguistic event previously uncovered in the affective computing community.

In the analysis of the speech utterances, instead of a large set of features we propose to extract very small, reduced-size feature sets. Specifically, we tried two state-of-the-art minimalistic sets (GeMAPS and extended GeMAPS) in addition to own proposals of spectral features in a highly informative frequency range: the mean log-spectrum of the speech signal and the mean log-spectrum of the signal at the auditory level.

A number of experiments were carried out with combinations of these sets. The results showed promising results using some of these combinations, in particular the extended GeMAPS plus our spectral features, which performed better than the baseline in the training data. Despite the performance drop for test evaluation, we showed that using less than 3% of the size in the extracted features we can obtain similar (or even better performances, such as the correlation on the training set) than the baseline trained with a very large number of features.

Future works will be devoted to research further the potential and benefits of these new sets of features with other classification/regression schemes, for example, dynamic models or deep neural networks that exploit also the local variability along the utterances (temporal profile of the paralinguistic phenomena). Also, a more in-depth analysis of the information provided at the auditory level could lead to the proposal of new features, extending the benefits that MLSa showed here on the training set.

ACKNOWLEDGEMENTS

The authors wish to thank to *Universidad Nacional de Litoral* (with PACT 2011 #58 and CAI+D 2011 #58-511) and *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina, for their support.

REFERENCES

[1] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, “The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language,” in *Proc. Interspeech 2016, ISCA, San Francisco, USA*, 2016.

[2] P. Rockwell, “Lower, slower, louder: Vocal cues of sarcasm,” *Journal of Psycholinguistic Research*, vol. 29, no. 5, pp. 483–495, 2000.

[3] G. A. Bryant and J. E. Fox Tree, "Recognizing verbal irony in spontaneous speech," *Metaphor and symbol*, vol. 17, no. 2, pp. 99–119, 2002.

[4] S. Peters, K. Wilson, T. Boiteau, C. Gelormini-Lezama, and A. Almor, "Do you hear it now? a native advantage for sarcasm processing," *Bilingualism: Language and Cognition*, vol. 19, no. 2, pp. 400–414, 2016.

[5] H. Lœvenbruck, M. A. B. Jannet, M. D'Imperio, M. Spini, and M. Champagne-Lavau, "Prosodic cues of sarcastic speech in french: slower, higher, wider," in *Proc. Interspeech 2013, ISCA, Lyon, France*, pp. 3537-3541, 2013.

[6] D. Voyer, S.-H. Thibodeau, and B. J. Delong, "Context, contrast, and tone of voice in auditory sarcasm perception," *Journal of Psycholinguistic Research*, vol. 45, no. 1, pp. 29–53, 2016.

[7] J. Tepperman, D. R. Traum, and S. Narayanan, " "Yeah right": sarcasm recognition for spoken dialogue systems." in *Proc. Interspeech 2006, ISCA, Pennsylvania, USA*, 2006.

[8] R. Rakov and A. Rosenberg, " "Sure, I did the right thing": a system for sarcasm detection in speech." in *Proc. Interspeech 2013, ISCA, Lyon, France*, pp. 842-846, 2013.

[9] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015, open access. [Online]. Available: http://doc.utwente.nl/98965/

[10] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, vol. 25, no. 3, pp. 556–570, 2011.

[11] S. A. Shamma, R. S. Chadwick, W. J. Wilbur, K. A. Morrish, and J. Rinzel, "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," *The Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 133–145, 1986.

[12] C. Martínez, J. Goddard, D. Milone, and H. Rufiner, "Bioinspired sparse spectro-temporal representation of speech for robust classification," *Computer Speech & Language*, vol. 26, no. 5, pp. 336–348, 2012.

[13] C. Martínez, J. Goddard, L. Di Persia, D. Milone, and H. Rufiner, "Denoising sound signals in a bioinspired non-negative spectro-temporal domain," *Digital Signal Processing*, vol. 38, pp. 22–31, 2015.

[14] E. Albornoz and D. Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, 2016.

[15] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *21st ACM International Conference on Multimedia*, Barcelona, Spain, October 2013, pp. 835–838.

[16] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in Emotion Science*, vol. 4, no. 292, pp. 1–12, 2013.

[17] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Feature extraction based on bio-inspired model for robust emotion recognition," *Soft Computing*, pp. 1–14, 2016.

[18] S. M. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature neuroscience*, vol. 8, no. 10, pp. 1371–1379, 2005.

[19] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, march 1992.

[20] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[21] N. Mesgarani and S. Shamma, "Denoising in the domain of spectrotemporal modulations," *EURASIP J. Audio Speech Music Processing*, vol. 2007, no. 3, pp. 1–8, Jul. 2007.

[22] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[23] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods*, pp. 185–208, 1999.