INTERSPEECH
2009 BRIGHTON

Recognising Interest in Conversational Speech – Comparing Bag of Frames and Supra-segmental Features

Björn Schuller and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
80333 München, Germany
Schuller@tum.de

10.21437/Interspeech.2009-484

Abstract

It is common knowledge that affective and emotion-related states are acoustically well modelled on a supra-segmental level. Nonetheless successes are reported for frame-level processing either by means of dynamic classification or multi-instance learning techniques. In this work a quantitative feature-type-wise comparison between frame-level and supra-segmental analysis is carried out for the recognition of interest in human conversational speech. To shed light on the respective differences the same classifier, namely Support-Vector-Machines, is used in both cases: once by clustering a 'bag of frames' of unknown sequence length employing Multi-Instance Learning techniques, and once by statistical functional application for the projection of the time series onto a static feature vector. As database serves the Audiovisual Interest Corpus of naturalistic interest.

Index Terms: interest recognition, multi-instance learning, feature relevance

1. Introduction

Opposing the majority of works in the field of speech signal analysis in search of emotion-related states that operate on a supra-segmental level, ever since works also operate directly on the frame-level [1, 2, 3, 4, 5]. In fact, certain application fields may require online incremental processing [6], and several other modalities are often processed on this level, e. g. facial expression and body gesture analysis, as well as neuro impulses and general bio-signals. Thus, (incremental) frame-level speech processing allows for easy integration with these modalities, potentially even in an early fusion.

Usually frame-level analysis of speech in search of affective cues is carried out employing only cepstral features and energy by Hidden Markov Models. This limitation of feature type is mostly stemming from the tools employed, which are usually designed for speech rather than emotion recognition. Few direct comparison exist [3, 7] that also consider further feature types employed in supra-segmental recognition of emotion-related states as pitch or Harmonics-to-Noise ratio: there it is shown that the supra-segmental modelling is superior. This is well in line with other studies (e. g. [8]) that considered only cepstral and energy feature information. Interestingly, also a gain by combination of frame-level and supra-segmental modelling could be demonstrated [8]. Yet, to the best knowledge of the authors, no study so far carried out a quantitative type-by-type direct comparison of these two.

To this aim Multi-Instance (MI) Learning techniques are applied in this work to classify a sequence of frames of unknown length on the frame-level by the exact same classifier as a single projected vector on the supra-segmental level. In principle, such a comparison could be made using Gaussian Mixture Models as done in [3] or any majority voting scheme. However, by that the frame-level information is not processed in one pass, i. e. only one frame is seen at a time. In contrast, the supra-segmental analysis profits from features derived from a whole emotionally or syntactically meaningful unit – usually words, chunks [7], or speech turns [1, 4, 5, 9]. This is overcome herein by processing a 'bag of frames' in one pass with one of the most popular classifiers in the field – a Support Vector Machine in an MI variant [10].

The data analysed is human-to-human conversational speech. The considered affective cues are three levels of interest reaching from bored to highly interested which bears great potential in many applications as customer interest detection, student tutoring systems or automatic meeting analysis [11]. While a fusion of further multiple streams as linguistic analysis of the spoken content, eye activity, and facial expression and contextual knowledge may help improve on this task [12], this work focuses on the so far strongest stream in an automatic detection scenario: the acoustic feature information.

Throughout the remainder of this paper the features on frame and supra-segmental level are introduced in sec. 2, the required classification paradigms in sec. 3, the data used in the evaluation in sec. 4, before presenting results and concluding in sec. 5, and sec. 6.

2. Frame-level and Supra-segmental Features

The basis is a set of 37 typical acoustic Low-Level-Descriptors well known to carry information about paralinguistic effects shown in Table 1. The features cover the common types pitch, energy, formants, cepstral, and voice quality:

Energy: these features model intensity, based on the amplitude in different intervals, with different weighting and transformation.

Pitch: this is the acoustic equivalent to the perceptual unit pitch. It is measured in Hz and bases on the autocorrelation function.

Formants: formants (i. e. spectral maxima) are known to model spoken content, especially lower ones. Higher ones however also represent speaker characteristics. Each one is fully represented by its position, amplitude and bandwidth.

Cepstral: Mel Frequency Cepstral Coefficients (MFCC) features (homomorphic transform with equidistant band-pass-filters on the Mel-scale) tend to strongly depend on the spoken content. Yet, they have been proven highly beneficial in practically any speech and most audio processing tasks.

Voice quality: jitter and shimmer are micro-perturbations based on pitch and intensity reflecting voice quality. As further low-level voice quality feature Harmonics-to-Noise Ratio (HNR) is added.

In order to calculate low-level descriptors, first the speech signal is transformed to 16 kHz, 16 bit. In general, a Hamming window function is used, except for the calculation of F0 and HNR where a Hanning window has been chosen. 100 fps are used with semi-overlapping windows. Energy resembles simple log frame energy. F0 and HNR calculation base on autocorrelation in the time domain with window correction. Formants base on 18-point LPC with root-solving and a pre-emphasis factor $\alpha = 0.7$. F0 and formant trajectories are globally optimized by use of dynamic programming. Low-level descriptors are smoothed by according techniques as semi-tone-interval filters or simple moving average low-pass-filtering to overcome noise. As a next step delta coefficients are added for each low-level descriptor.

A strictly systematic generation of features was chosen for the construction of a large feature space as basis for the supra-segmental analysis. Such an approach generally leads to >1k features.

Following the typical static classification strategy used in emotion recognition, next a total of 19 statistical functionals is applied to each of the 2 x 37 low-level descriptors. The obtained multivariate time series of variable length is projected onto a single 1 406 dimensional feature vector. Here again it is decided for a typical selection of common functionals covering the first four statistical moments, quartiles, extremes, ranges, positions, and zero-crossings as depicted in as shown in Table 2.

The three position related functionals lead to a sub-group of features with the physical unit of ms which are often treated as duration features, though having a number of diverse low-level descriptors as basis. It is refrained from inclusion of further duration related features such as those based on, e. g. lengths of pauses or syllables because this information cannot easily be integrated in the strictly systematic generation approach: it is modelled in a general value series rather than in a time series. Also, such information requires additional higher level analysis and would thus violate a fair comparison with the frame-level processing.

Table 1: *Low-Level Descriptors used.*

37 Low-Level Descriptors
Pitch (F0), Frame Energy, Envelope
Mel-Frequency Cepstral Coefficients (MFCC) 1-16
Formant 1-5: Amplitude, Bandwidth and Position
Shimmer, Jitter
Harmonics-to-Noise Ratio (HNR)

Note that this exact feature set was has proven itself in a number of emotion [8] and further emotion-related classification tasks as intimacy [9] and the aimed at interest recognition [12, 11].

3. Single and Multi-Instance Learning

In Multi-Instance Learning a 'bag of instances' is labelled by the same label. Usually the idea behind is that the labels of individuals withing the bag are not known and may differ. However, one instance being of the class of the bag is sufficient for it to be labelled, accordingly. There is some practical consideration why this may be adequate to model emotion in speech: as shown in [13], emotional turns usually contain a high percentage of neutral speech: on average among emotions 42.5 % of the words in emotional turns were shown to be neutral. Thus, a bag (e. g. all frames of a speech turn) labelled as angry may also contain neutral instances (e. g. frames) without 'pollution' of the training material. Further, operating on a frame-level, multiple instances will be short silences or potentially dominant background noises rather than speech of the target speaker, thus also not portraying the target emotion. In this respect Multi-Instance Learning seems generally well suited to model emotion in a detection scenario.

Table 2: *Functionals applied to Low-Level Descriptor contours used for systematic construction of a 1 406 dimensional acoustic feature space on the supra-segmental level.*

19 Functionals	
Mean, Centroid, Std. Dev.	Quartiles 1, 2, 3
Skewness, Kurtosis	Quartile 1 - Minimum
Zero-Crossing-Rate	Quartile 2 - Quartile 1
Max/Min Value, Range	Quartile 3 - Quartile 2
Relative Max/Min Position	Maximum - Quartile 3
95 % Roll-Off-Point	

However, in this work it is primarily employed to provide equal testing conditions when comparing frame-level and supra-segmental level modelling of emotional speech by using the same classification principle: as in [14] each frame belonging to a sub-speaker turn is assigned to the same bag. This bag is labelled with the sub-speaker turn label – just as the feature vector obtained by statistical functional based projection onto a single vector. This allows to classify a sequence of unknown length and a single vector with the same basic classifier – in this work Support Vector Machines (SVM) in a Multi-Instance implementation (MI-SVM) [15]. For a detailed description of MI-SVM the reader is referred to [10].

4. Audiovisual Interest Corpus

In the scenario setup, an experimenter and a subject are sitting on both sides of a desk. The experimenter plays the role of a product presenter and leads the subject through a commercial presentation. The subject's role is to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter considering his/her interest to the addressed topics without respect to politeness. Visual and voice data is recorded by a camera and two microphones, one headset and one far-field microphone.

After the final recording the Audiovisual Interest Corpus (AVIC) shows the following parameters and statistical figures for audio: audio sampling rate: 44.1 kHz, audio quantisation: 16 Bit, left audio channel: lapel microphone, right audio channel: far-field microphone. 21 subjects (10 of them female) took part, three of them Asian, the others European. The language throughout experiments is English, and all subjects are very experienced English speakers. Three age categories were defined during the specification phase (<30 years, 30-40 years, >40 years) for balancing. The mean age of male subjects resembles 32.7 years, of female subjects accordingly 30.1 years. The total recording time for males resembles 5:14:30 h, for females 5:08:00 h. By age categories the recording times are 4:40:40 h for <30 years, 4:10:20 h for 30-40 years, 1:31:30 h for >40 years. Likewise, a total of 10:22:30 h was recorded.

To acquire reliable labels of a subject's 'Level of Interest' (LOI) as detailed in the ongoing, the entire material was segmented in speaker and sub-speaker turns and subsequently labelled by 4 male annotators, independently. The LOI is annotated for every sub-speaker turn.

Five LOI were distinguished in the first place: LOI-2 – *Disinterest* (subject is bored listening and talking about the topic, very passive, does not follow the discourse), LOI-1 – *Indifference* (subject is passive, does not give much feedback to the experimenter's explanations, unmotivated questions if any), LOI0 – *Neutrality* (subject follows and participates in the discourse, it can not be recognised, if she/he is interested or indifferent in the topic), LOI1 – *Interest* (subject wants to discuss the topic, closely follows the explanations, asks some questions), LOI2 – *Curiosity* (strong wish of the subject to talk and learn more about the topic).

For automatic processing a fusion of these LOI to a 'master LOI' was automatically fulfilled by the following scheme of different cases of Inter Labeller Agreement (ILA) and confidence bounds:

- Same rating by all annotators: ILA 100 %;
  Master LOI := LOI of majority
- Same rating by 3 of 4 annotators: ILA 75 %;
  Master LOI := LOI of majority
- Same rating by 2 annotators: ILA 50 %
  > If other 2 annotators agree:
  Master LOI := '?' (undefined)
  > If other 2 annotators disagree:
  Master LOI := median LOI.
  In this case an additional confidence measure C is derived from the standard deviation σ of the LOI over all annotators: $C = 1 - 0.5 \cdot \sigma$.

The overall annotation contains sub-speaker and speaker turn segments in millisecond resolution, spoken content, non-verbals, individual annotator tracks, and Master LOI with confidence in XML-format provided by use of ANVIL. A sub-speaker turn is thereby defined by a turn lasting longer than 2 sec. Turns are split by punctuation and syntactical and grammatical rules until each segment lasts shorter than 2 sec.

The database comprises 12 839 sub-speaker turns. There is a total of 18 581 spoken words, and 23 084 word-like units including non-linguistic vocalisations (19.5%). A very low κ-value of $\kappa = 0.09$ and standard deviation for the LOI of the labellers of $\sigma = 0.54$ is observed for the database at this point. The ILA is therefore pruned of undefined sub-speaker turns (labelled with '?') and sub-speaker turns of the medium LOI with a confidence $C < 1.0$. Through this reduction of sub-speaker turns, the agreement of the four annotators increases to a substantial κ-value of $\kappa = 0.62$ with $\sigma = 0.23$, and the distribution of the instances over the LOI is more balanced.

As too few items for LOI-2 and LOI-1 have been seen, these were clustered together with LOI0, resulting in an LOI scale of LOI0 to LOI2. Thereby final values of $\kappa = 0.66$ with $\sigma = 0.20$ are observed. The exact LOI-distribution of the single labellers and ILA in this reduced set of sections is as depicted in Table 3.

For further details on the corpus as the annotation workflow, inter-labeller kappa-values, original distribution and mapping onto a regression problem the reader is referred to [12].

Table 3: *Distribution of the Level of Interest (LOI) over various labellers (Lab.) after the rejection of diffuse sub-speaker turns (SST). Mean μ and standard deviation σ of the LOI are also provided within the interval [0;2].*

#SST	**LOI0**	**LOI1**	**LOI2**	**μLOI**	**σLOI**
Lab. 1	257	569	170	0.91	0.65
Lab. 2	311	526	159	0.85	0.67
Lab. 3	175	643	178	1.00	0.60
Lab. 4	311	652	33	0.72	0.52
Male	150	306	64	0.83	0.62
Female	166	204	106	0.87	0.75
ILA	**316**	**510**	**170**	**0.85**	**0.68**

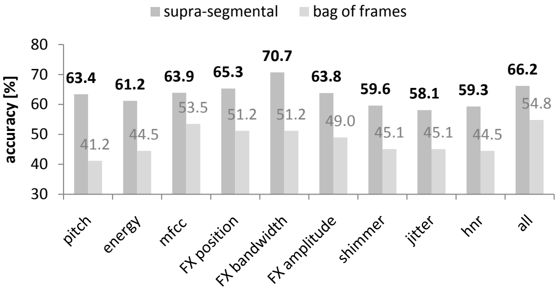

Figure 1: *Absolute accuracy in percent correct per feature type by supra-segmental (bold printed) and bag-of-frames modelling. SVM in 3-fold speaker-independent cross-validation. AVIC database.*

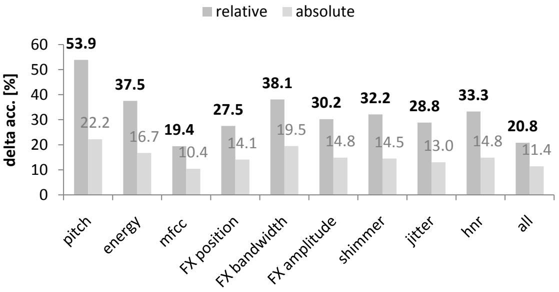

Figure 2: *Relative (bold printed) and absolute delta accuracy in percent per feature type, i. e. gain by supra-segmental over bag-of-frames modelling. SVM in 3-fold speaker-independent cross-validation. AVIC database.*

5. Experimental Comparison

In the following the results on the AVIC corpus for the analysis of interest in speech by either frame-level or supra-segmental modelling are presented. As evaluation strategy a speaker-independent 3-fold cross-validation is employed. This is obtained by partitioning the database into three speaker groups, each in age and gender balance, as carried out in [11].

Figure 1 visualises the respectively obtained mean accuracies in percent correctly assigned sub-speaker turns per feature group for the bag of frames by MI-SVM in comparison to the supra-segmental modelling by SVM. Polynomial kernels are each used. Multi-class discrimination is obtained by pairwise one-vs.-one decisions. Learning is carried out by Sequential Minimal Optimisation.

As can be seen from these figures, spectral information prevails in both cases – the frame and the supra-segmental level. However, in supra-segmental analysis formants come first (bandwidth first, subsequently position and last amplitude), followed by cepstral information. This is reversed on the frame-level, where MFCC dominate. The combination of all feature types leads to an improvement only on the frame-level. This could be overcome in supra-segmental modelling by optimising the feature space by an adequate de-correlating search and target function, as done in [12]. However, in this work the focus lies on differences rather than on highest obtainable accuracy obtained by optimisation with repeated measurement on the same dataset.

Comparing accuracies on frame and supra-segmental level, it is obvious that the latter is to be preferred in any case. Still, differences exist with respect to the individual gap concerning the feature type. Figure 2 depicts this differences in accuracy by absolute and relative gain of supra-segmental over bag of frames accuracy. Obviously, pitch profits most from supra-segmental representation (>50 % relative improvement) – just as one would assume. Least profit is observed for cepstral features (<20 % ralative improvement) – again well meeting expectation. The other feature types are found around 30 % relative improvement. Interstingly, also the combination off all feature types shows lower benefit. However, this may partly stem from increased vector dimension (cf. above).

6. Conclusion

In this work the differences between frame-level and supra-segmental features were quantitatively revealed for speaker-independent recognition of three levels of naturalistic interest in conversational human speech. It was shown that pitch, the bandwidth of formants, and energy profit most from supra-segmental analysis, while MFCC suffer less from frame-level modelling than any other considered type. Further, the relevance of feature groups was analysed, whereby spectral and cepstral come first, followed by prosodic and least voice quality features. However, more elaborate voice quality features [16] may change this picture. Also, Support Vector Machines were used throughout. While the whole context of frames was provided to the SVM, other architectures as the Long-Short-Term-Memory networks introduced in [17] that learn the optimal context size or Hidden Markov Models and general Dynamic Bayesian Networks with warping ability may lead to different findings on frame-level.

While Multi-Instance Learning served mostly for better comparability in this work, future ambitions might truly benefit from its power to allow for corrupted bags with respect to neutral or differently emotional speech. However, this will probably rather be beneficial on the word level, i. e. 'bags of words', but in an acoustic feature vector sense with frame series projected onto one vector per word.

Further, as stated, the ability to add new frames to existing bags may well be utilised to allow for incremental processing. Here, it will be interesting to analyse the improvement of an estimate for a larger unit as a speech turn with an increasing amount of frames available.

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

8. References

[1] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Proceedings of the EUROSPEECH 2001*, 2001, pp. 2267–2270.

[2] T. Nwe, S. Foo, and L. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, pp. 603–623, 2003.

[3] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Proceedings ICASSP*, vol. 2, Hong Kong, China, 2003, pp. 1–4.

[4] Z. Inanoglu and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," in *Proceedings of the 10th International Conference on Intelligent User Interfaces*, San Diego, California, USA, 2005, pp. 251–253.

[5] J. Wagner, T. Vogt, and André, "A Systematic Comparison of different HMM designs for Emotion Recognition from Acted and Spontaneous Speech," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds. Berlin-Heidelberg: Springer, 2007, pp. 114–125.

[6] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proceedings 4th Intern. Workshop on Human-Computer Conversation*, Bellagio, 2008.

[7] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of the Interspeech*, Brighton, UK, 2009.

[8] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," in *Proceedings Interspeech*, Antwerp, 2007, pp. 2249–2252.

[9] A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl, "Mothers, Adults, Children, Pets — Towards the Acoustics of Intimacy," in *Proceedings of ICASSP 2008*, Las Vegas, 2008, pp. 4497–4500.

[10] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances of Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, 2009, pp. 561–568, cambridge.

[11] B. Schuller, N. Köhler, R. Müller, and G. Rigoll, "Recognition of Interest in Human Conversational Speech," in *Proceedings Interspeech*, Pittsburgh, 2006, pp. 793–796.

[12] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, p. 17 pages, 2009, to appear.

[13] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Does Affect Affect Automatic Recognition of Children's Speech?" in *Proceedings of the 1st Workshop on Child, Computer and Interaction*, Chania, Greece, 2008.

[14] M. Shami and W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-corpus Study," in *Speaker Classification II*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Heidelberg - Berlin - New York: Springer, 2007, vol. 4441, pp. 43–56.

[15] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques, 2nd Edition*. San Francisco: Morgan Kaufmann, 2005.

[16] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[17] M. Wllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008*.