ACOUSTIC MEASURES CHARACTERIZING ANGER ACROSS CORPORA COLLECTED IN ARTIFICIAL OR NATURAL CONTEXT

Marie Tahon, Laurence Devillers

{mtahon,devil}@limsi.fr

ABSTRACT

This paper aims at studying differences between acoustic manifestations of anger across corpora collected in artificial, manipulated or natural context. It aims further at finding measures of naturalness in emotive corpora. Evaluating the degree of naturalness of a corpus can be challenging unless given knowledge upon the task. In corpora consisting of rather acted data, anger is often stronger; we believe that a kind of distance can be computed between anger and the overall corpus data. Such a distance is introduced in this work and evaluated with state-of-art acoustic descriptors in 3 collected corpora. We show the observed differences between the acoustic features obtained with anger samples in these different contexts and propose measures of naturalness.

Index Terms — Emotion-detection, anger, naturalness, prototypical emotion.

1. INTRODUCTION

Emotion detection across different corpora and combination of several corpora in order to build emotion detection systems are actual challenges. This paper aims at studying differences between acoustic manifestations of anger across corpora and also at finding measures of naturalness in emotive corpora. Task, sound quality, how the speaker is recorded, speakers' characteristics (age, sex, voice quality) have a great influence on the decoder results. Every corpus used in emotion detection is different one from another. The more the task is definite, the more the detection would be efficient. As a result, detection scores with acted corpus are much better than with spontaneous corpus [8]. But actual challenges tend to focus on real-life corpora. To compare emotions across corpora we face the fact that expression of the same emotion vary according to the context (from real-life to acted). In real-life context, social control tends to shaded and mixed emotions [1], whereas in acted context, emotions tend to be prototypical.

Anger is the emotion which is usually found among best recognized within automatic decoding in a given corpus, and there is anger in most of the corpus. Thus, we decided for strong and non-ambiguous anger as basis for comparison in three corpora collected in different contexts. We will compare acoustic results computed on anger samples across the three corpora.

In this study, we will deal with the breakdown into acted/non-acted databases and we will not tackle the question of the natural dimension defined by Scherer [7] We would like to estimate how a corpus is naturalistic/induced using acoustic measures of anger as a reference. We would like to have a mixture of measures to estimate the corpus position on the naturalistic/induced scale. First we describe the three corpora we are working on. In a second part, we define the acoustic features studied in this paper. The third part consists in the feature analysis and finally we conclude by possible improvements and outlooks.

2. DATABASES

Our study is based on three corpora: CINEMO is an acted corpus with oriented context, so called “realistic corpus”, JEMO is a prototypical portrayed emotion corpus and CEMO is a natural corpus collected in a medical emergency call-center.

The CINEMO database [7] is composed of 29 selected excerpts from 12 French movies. 51 speakers (21 female, 30 male) were told to dub or improvise (with and without subtitles) 2 or 3 characters from different movies. Our data set consists in 3992 segmented samples, extracted from the dubbed and improvised instances after applying annotation agreement of two coders. 16 labels were annotated. These labels are part of 5 macro-classes (joy/ sadness/ fear/ anger/ neutral). In our experiment we will work with two corpora: CINEMO-all (all samples) and CINEMO-anger, a sub-corpus with only strong and non-ambiguous anger.

The JEMO corpus has been collected within the framework of an emotion detection game: people tends to play basic emotions for being recognized by the computer. The game is able to detect anger, joy, sadness, anger or “neutral” in the player's voice (content

of speech is free). We will use the whole JEMO corpora (called JEMO-all) and a sub-corpus of JEMO composed of strong and non-ambiguous angry samples. Emotions in JEMO are stereotypical emotions.

CEMO is a collection of spontaneous emotions recorded in an emergency call-center [1]. They have been collected via phone. For our experiment we will use a small part of this corpus called CEMO-all. This part is composed of anger and relief speech samples. The sub-corpus CEMO-anger is composed only of strong and non-ambiguous anger samples.

Each sub-corpus Anger is composed of the strongest instances of anger, chosen manually. In the table below (Table I), the number of instances and the number of speakers of the six corpora used in this study are classified according to the speaker's sex.

TABLE I: DATA

#Seg. (#speakers)	Male	Female
CINEMO-all	2210 (29)	1379 (21)
CINEMO-anger	140 (11)	191 (11)
JEMO-all	1468 (18)	1142 (17)
JEMO-anger	86 (16)	41 (11)
CEMO-all	133 (55)	133 (80)
CEMO-anger	98 (26)	71 (34)

3. ACOUSTIC FEATURES AND COMPUTING

3.1. Acoustic features

As the vocal production system (size of the vocal track, vocal resonances) is different between male and female, a lot of spectral parameters will be different. Thus, we are going to study both sex separately [3]. Music features appears to be interesting for emotion detection. The features marked with a star (*) in Table II, come from music description [2, 5].

TABLE II: ACOUSTIC FEATURES

Feature type	Name
Spectral	Rolloff (5%*, 25%, 50%, 75%, 95%*), centroid, slope* (2), BarkBandEnergy (1 to 24), BandEnergy* (4)
Cepstral	MFCC 0 to 12
Energy	MeanLoudness, RMSenergy
Time-domain	ZCR, VoicedRatio,
Voice	RatioF0, varF0, F2-F1, F3-F2, varF1, varF2, varF3, HNR

3.2. Computing

Time-window on which is computed the feature, normalization to the speaker and sound quality require a great attention. These aspects will not be studied here, but they are fundamental. The larger time-window is the voiced segment. According to Martinet [4]. vowels contain more information (linguistic and paralinguistic) than the rest of the speech signal (consonant, noise, etc.). For this first experiment, we only compute acoustic features on voiced segments.

PRAAT gives us fundamental frequency, formants and micro-prosodic features (HNR, VoicedRatio) on voiced segments. With MATLAB, we compute also ZCR, and energy features (MeanLoudness and RMSenergy). Except micro-prosodic features, we compute mean values and variances of each time variables (F0, formants, and energy) on the whole voiced segment. Then a Fast Fourier Transform on the whole voiced signal gives us all spectral and cesptral features. Of course, if there is no voiced part in the segment, it is impossible to compute any feature. We will also not use small segments (duration lower than 50ms).

4. RESULTS AND ANALYSIS

4.1. Expert investigation on acoustic features

Each Table presents the mean values and the standard deviation (variances) for the main acoustic feature for female samples. The first important result is that, generally speaking, feature means are higher for Corpus-anger than for Corpus-all. The standard deviation values have been computed in order to select the most relevant features. Variances is the standard deviation normalized by the mean. The general results tend to be the same with male samples.

Energy features:

TABLE III: ENERGY FEATURES

	CINEMO		JEMO		CEMO	
	Anger	All	Anger	All	Anger	All
mean values on overall corpus						
MeanLoudnes	4.2E+1	3.3E+1	5.5E+1	3.3E+1	4.4E+1	4.1E+1
RMSenergy	3.0E-4	1.0E-4	3.0E-4	2.0E-4	4.0E-4	3.0E-4
relatives variances on overall corpus						
MeanLoudnes	0.48	0.53	0.63	0.66	0.29	0.34
RMSenergy	2.75	3.04	2.94	3.6	0.88	0.93

Activation is the main dimension for recognizing anger from other emotions. A lot of detection systems defined activation with energy features such as MeanEnergy. In Table III, we can see that both Loudness and RMSenergy are higher for anger corpus that for the all corpus. A comparison of variances values

on both MeanLoundess and RMSenergy, shows that the RMSenergy varies much more than loudness (see Table III). So MeanLoudness seems to be more suitable for emotion detection than RMSenergy.

Spectral features:

TABLE IV: SPECTRAL FEATURES

	CINEMO		JEMO		CEMO	
	Anger	All	Anger	All	Anger	All
mean values on overall corpus						
RollOff5%	2.4E+2	2.2E+2	3.2E+2	2.2E+2	3.4E+2	3.6E+2
RollOff50%	4.6E+2	4.3E+2	8.3E+2	5.1E+2	6.6E+2	7.0E+2
RollOff95%	1.4E+3	1.4E+3	2.4E+3	1.7E+3	2.1E+3	2.0E+3
centroid	1.6E+3	1.5E+3	1.9E+3	1.5E+3	1.3E+3	1.3E+3
slope	-1.0E-3	-7.0E-4	-1.5E-3	-7.0E-4	-2.5E-3	-2.1E-3
[0-250Hz]	2.4E+2	1.7E+2	1.7E+2	7.2E+1	2.5E+1	2.0E+1
[250-650Hz]	8.0E+2	5.2E+2	1.6E+3	4.1E+2	4.4E+2	3.5E+2
[650-1kHz]	4.8E+2	2.3E+2	2.0E+3	5.1E+2	2.3E+2	1.9E+2
Bark 1-7	6.0E+0	5.3E+0	7.0E+0	4.8E+0	6.0E+0	5.2E+0
Bark 8-14	4.3E+0	3.5E+0	7.6E+0	3.9E+0	6.5E+0	5.7E+0
relatives variances on overall corpus						
RollOff5%	0.42	0.80	0.59	0.92	0.22	0.56
RollOff50%	0.49	0.91	0.84	1.00	0.33	0.50
RollOff95%	0.63	0.81	0.48	0.73	0.29	0.31
centroid	**0.27**	**0.33**	**0.25**	**0.36**	**0.17**	**0.20**
slope	0.97	1.21	1.40	1.53	0.69	0.71
[0-250Hz]	2.19	2.33	3.64	2.79	2.58	2.35
[250-650Hz]	1.92	2.62	5.86	3.31	1.18	1.15
[650-1kHz]	2.98	3.80	7.60	5.47	1.25	1.21
Bark 1-7	0.59	0.67	0.78	0.74	0.53	0.55
Bark 8-14	0.68	0.77	1.11	1.00	0.52	0.55

Spectral values are often higher for anger than for the other emotional classes as shown in table IV in CINEMO and JEMO corpora, the trend is attenuated in CEMO. This is true for all Bark Bands, Rolloff frequencies, centroid, slope and Band Energy. As we did not normalized features to the time-energy variable, we cannot prove that spectral energy values are linked with time-energy. Spectral features standard deviations are quite often below 1 (grey spaces) except for Band Energy. Both Bark Bands (Bark scale) and Band Energy (Hertz scale) aim at giving the amount of energy in a defined frequency band but variance with Band Energy are higher than with Bark bands (perceptual features). The lowest deviation is for the centroid feature.

TABLE V: CEPSTRAL FEATURES

	CINEMO		JEMO		CEMO	
	Anger	All	Anger	All	Anger	All
mean values on overall corpus						
MFCC0	1.1E+1	1.0E+1	1.3E+1	1.0E+1	1.3E+1	1.2E+1
MFCC1-7	4.4E-1	4.4E-1	5.8E-1	4.9E-1	5.8E-1	5.8E-1
relatives variances on overall corpus						
MFCC0	0.22	0.30	0.24	0.35	0.22	0.27
MFCC1-7	0.72	0.71	0.82	0.77	0.61	0.63

Cesptral features:

MFCC means, except for MFCC0 that corresponds to a measure of the signal energy based on Mel scale, do not vary with the corpus, neither with anger. As the variances do not vary with anger, we can suppose that they are much linked to speakers than to emotions. Table V shows only the first 7 MFCC, but the results are the same for the other 5 coefficients.

Time domain features:

When looking to ZCR, HNR and VoicedRatio anger means are not always stronger than all means. To our point of view, the use of ZCR is conditioned by phoneme segmentation (vowels, voiced, unvoiced consonants) and by recording conditions. As a result, it cannot be an interesting feature for emotion detection without any alignment with the phonetic content of the speech.

TABLE VI: TIME DOMAIN AND VOICE FEATURES

	CINEMO		JEMO		CEMO	
	Anger	All	Anger	All	Anger	All
mean values on overall corpus						
RatioF0	1.2E+0	1.3E+0	1.3E+0	1.3E+0	1.3E+0	1.3E+0
VarF0	1.7E+1	1.8E+1	2.5E+1	2.3E+1	2.0E+1	1.9E+1
F2-F1	2.2E+3	2.1E+3	2.1E+3	1.9E+3	1.5E+3	1.4E+3
F3-F2	-7.4E+2	-9.3E+2	6.2E+1	-4.4E+2	1.3E+3	1.2E+3
varF1	1.3E+2	1.3E+2	1.3E+2	1.2E+2	7.6E+1	7.3E+1
varF2	2.8E+2	2.5E+2	2.5E+2	2.0E+2	1.4E+2	1.3E+2
varF3	1.6E+3	1.5E+3	1.6E+3	1.1E+3	1.1E+2	9.8E+1
VoicedRatio	4.7E-1	4.4E-1	3.2E-1	4.1E-1	3.2E-1	3.4E-1
HNR	1.1E+1	1.2E+1	1.1E+1	1.1E+1	1.0E+1	1.0E+1
ZCR	2.4E+2	2.6E+2	5.9E+2	3.8E+2	6.1E+2	5.2E+2
relatives variances on overall corpus						
RatioF0	0.18	0.25	0.21	0.31	0.32	0.33
VarF0	0.75	0.98	0.78	1.20	1.06	1.07
F2-F1	0.16	0.30	0.12	0.39	0.14	0.27
F3-F2	2.12	1.36	12.48	3.32	0.15	0.27
varF1	1.07	1.11	0.77	1.23	0.66	0.72
varF2	0.64	0.74	0.62	0.93	0.55	0.61
varF3	0.49	0.57	0.52	0.82	1.38	1.32
VoicedRatio	0.32	0.31	0.31	0.45	0.36	0.40
HNR	0.17	0.22	0.30	0.37	0.19	0.19
ZCR	0.89	0.88	1.36	1.16	0.71	0.73

The HNR value is mainly used for signal quality. As variances are almost the same for a given corpus (see grey spaces), it seems that the quality vary with the corpus but not with emotion. VoicedRatio is probably highly linked with the segmentation protocol and the fundamental frequency detection errors of PRAAT.

Voice features:

Because the difference F3-F2 has an important variation (table VI, blue spaces), it appears that it is not suitable for our emotion detection. If we do not take into account the variance, all harmonic features, behave according to the general trend (red spaces). The problem

with the study of formants is that we cannot estimate the phoneme dependency. Moreover, when we use directly the F0 time variable, we do not know how many detection errors PRAAT have done. That is the reason why results on voice features must be balanced.

4.2. Acoustic features and naturalness

Now we have studied acoustic features, and their variations with anger, we would like to measure a distance between anger and the rest of the corpus for most relevant features. We have computed the relative difference between anger-mean and all-mean of each remaining features. This relative difference is called distance. If this distance is negative, then the value of the feature for anger is lower than for all the corpus. Table VII shows the 10 first acoustic features arranged from the higher distance to the lower. This classification aims at giving an order of the three corpora: CEMO-CINEMO-JEMO even if some distances are, sometimes, negative.

TABLE VII: DISTANCE BETWEEN ANGER AND ALL

	Anger over All ratio		
	CINEMO	JEMO	CEMO
[250-650Hz]	0.53	2.93	0.27
[650-1kHz]	1.07	2.81	0.23
[0-650Hz]	0.51	2.76	0.27
RMSenergy	2.00	0.50	0.33
Bark9	0.21	1.28	0.12
Bark16	0.28	1.24	0.13
[0-250Hz]	0.42	1.34	0.25
Bark18	0.25	1.19	0.13
Bark17	0.23	1.10	0.14
Bark15	0.26	1.08	0.13
slope	0.43	1.14	0.19
Bark19	0.23	1.05	0.13
Bark20	0.26	1.02	0.14
Bark13	0.28	0.94	0.15
Bark12	0.25	0.89	0.15

In conclusion, the best acoustic features for "measuring" the naturalness of a corpus are mainly spectral features. All the analyses we have done in this section were for female samples, male samples give similar results.

5. CONCLUSION AND OUTLOOKS

Are we able to compare strong anger across corpora? Some perceptual features like MeanLoudness, BarkBand Energy, centroid show a general trend for male and female samples: mean features are higher for Corpus-anger than for Corpus-all. It is possible to have a kind of classification of corpora based on a mix of acoustic features.

The more the corpus is acted, the more differences between anger and the entire corpus are accentuated. This paper shows that acted corpus contains emotions we are unable to detect in spontaneous corpus; as a result it seems important to work with both type of corpus. Taking strong and non-ambiguous anger as an emotion reference, the relative distance between acoustic mean values of Corpus-anger and Corpus-all gives a classification with some specific acoustic features. It could be interesting to compare our results based on strong anger with other emotions.

Characterizing the naturalness of the context of a corpus with acoustic features can allow more studies based on spontaneous corpora: a lot of work has been done with prototypical emotions in corpus data, this classification aims at studying more spontaneous corpora.

REFERENCES

[1] L. Devillers,L. Vidrascu, L. Lamel (2005), "Challenges in real-life emotion annotation and machine learning based detection", Journal of Neural Networks 2005, special issue: Emotion and Brain, vol 18, number 4, pp. 407-422.

[2] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpousis (2007), "The HUMAINE database: Adressing the collection of annotation of naturalistic and induced emotional data". In A. Paiva, R. Prada, R.W. Picard editors, *Affective computing and intelligent interaction*, pp 488-500, Berlin-Heidelberg. Springer.

[3] M. Garnier, D. Dubois, J. Poitevineau, N. Henrich, M. Castellengo, (2004) "Perception et acoustique dans la qualité vocale dans le chant lyrique", ICVPB.

[4] M. Kotti, C. Kotropoulos, (2008) "Gender classification in two Emotional Speech databases", International Conference on Pattern Recognition.

[5] A. Martinet, (1980) "Eléments de linguistique générale", ed. Armand Colin.

[6] G. Peeters, (2004) "A large set of audio features for sound descriptions (similarity and classification) in the CUIDADO project.

[7] N. Rollet, A. Delaborde, L. Devillers, (2009) "Protocol CINEMO: The use of fiction for collecting emotional data in naturalistic controlled oriented context", ACII.

[8] K.R. Scherer, T. Johnstone and G. Klasmeyer (2003), "Vocal expression of emotion", In R.J. Davidson, K.R. Scherer and H.H. Goldsmith editors, *Handbook of Affective Sciences*, chapter 23, pp 433-456. Oxford University Press, Oxford, New-York.

[9] S. Steidl, B. Schüller, A. Batliner, D. Seppi, "The hinterlands of emotions, facing the open-microphone challenge", ACII 2009.