Deep-learning for emotion recognition

Caroline ETIENNE

PhD Director : Pr. Laurence DEVILLERS & DreamQuark supervisor : Dr. Benoit SCHMAUCH

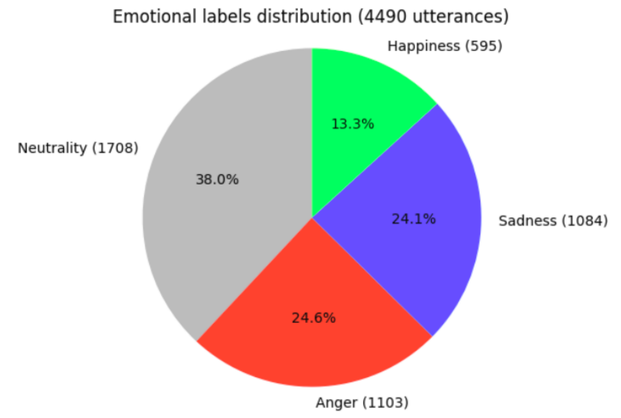
Data Science Summer School DS³ - École Polytechnique - 29 August 2017

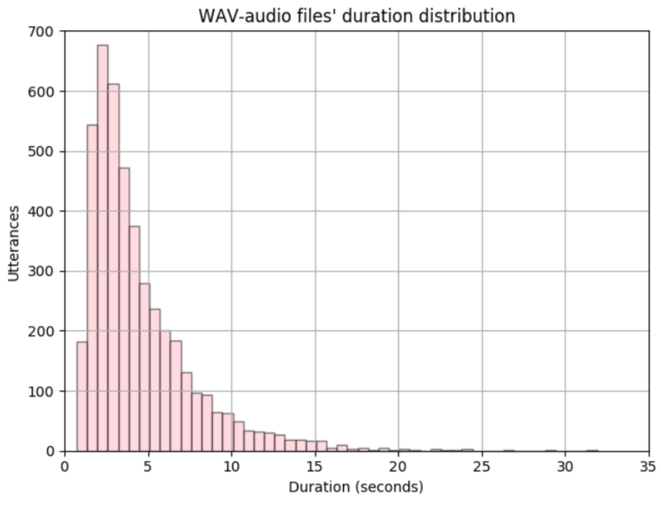
Introduction

This study is about speech emotion recognition systems. Speech emotion recognition with classification algorithms in human-machine interaction is a very challenging study.

Emotion give information about people and detecting them automatically can promote better communication. Today, call centers provide service to consumers and end users.
Determining the consumer satisfaction could give information, duration of call, and improving the companys image from the consumers perspective or determining the effectiveness of the call center employee.

Our current goal is to compare a study between an approach with selected acoustic features and an approach with low-level features such as spectrograms or time series.





Methods

Data description

The dataset used in this study is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database collected at the University of Southern California.
The dataset consists of around 12 hours of acted and spontaneous multimodal interactions from 10 human actors (5 male and 5 female). We only use audio files for our experiments.

1 utterance = 1 label (neutral state, sadness, angry, happiness)

The entire set for 4 emotions has 4490 improvisations and scripts audio files. Utterances have variable durations.

Pre-processing

- Accoustic-features extraction

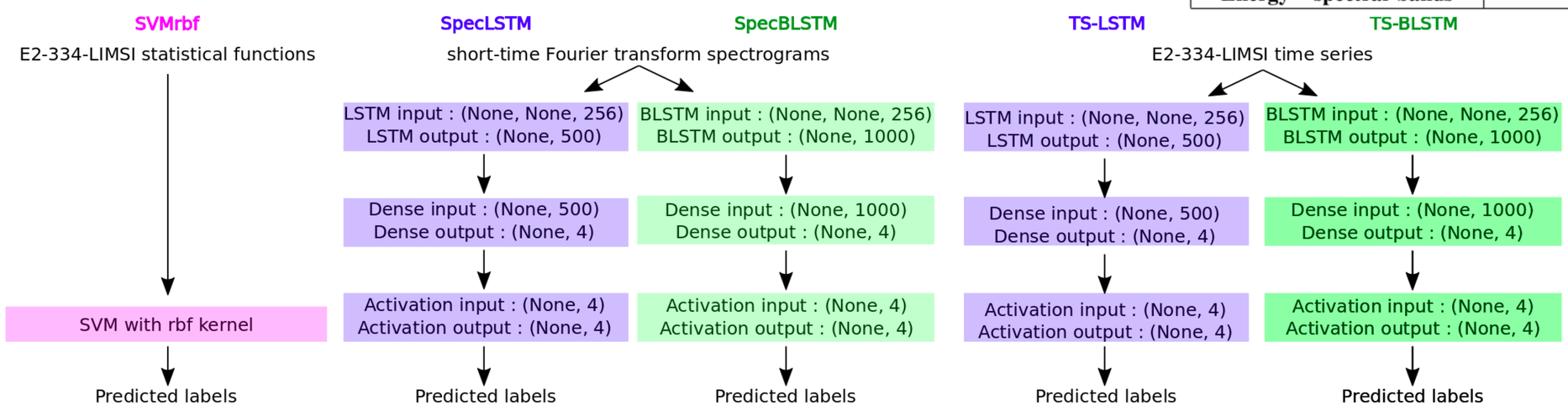
Acoustic descriptors used for emotion detection are borrowed from the fields of phonetics, speech recognition and music recognition and have been used to measure many phonation and articulation aspects.
These descriptors include acoustic parameters in the frequency domain (e.g. fundamental frequency F_0), in the amplitude domain (e.g. energy), in the time domain (e.g. rhythm) and in the spectral domain (e.g. spectral envelop or energy per spectral bands). Many studies show the interest of combining many of these parameters.
Temporal characteristics of the data are obtained with statistical functionals.
The set of descriptors used is E2-334-LIMSI.

Set of acoustic features (E2-334-LIMSI). 10 functionnals : min, max, mean, median, std, slope, centroid, spread, kurtosis, skewness. Yaafe parameters : blockSize=320, stepSize=160.

LLD	Functionals	Voiced	All
ZCR	10 func.	**0**	**10**
Jitter (absolute)	mean, std	2	
Jitter (relative)	mean, std	2	
Shimmer	mean, std	2	
Shimmer (dB)	mean, std	2	
Punvoiced			1
Voice Quality		**8**	**1**
F_0	mean, max, min, median, slope	12	
VoicedFrames			1
Number of Voice Break			1
Degree of Voice Break			1
Pitch		**12**	**3**
Roll Off 95%	10 func.		10
Spectral Decrease	10 func.		10
Spectral Variation	10 func.		10
Perceptual Spread	10 func.		10
Perceptual Sharpness	10 func.		10
Spectral Flatness	10 func.		10
Spectrum		**0**	**60**
Specific Loudness 0-21 (dB)	10 func.		240
Energy - spectral bands		**0**	**240**

- Spectrograms

A second approach is we experiment with features from the speech spectrogram. Spectrogram are computed with short-time Fourier transform (STFT) to analyze the frequency spectrum of signal. The Discrete-time Fourier transform is applied to a temporal window, or a frame.



Our five models

Results

Experiments scores. *%WA, weighted accuracy ; %UA, unweigthed accuracy.*

Model	%WA	%UA
E2-SVMrbf	51.06	49.65
SpecLSTM	47.0	46.45
SpecBLSTM	51.71	48.31
TS-LSTM	47.24	46.79
TS-BLSTM	48.44	48.10

Experiments computations times per fold and number of epochs.

Model	Time per fold	Training
E2-SVMrbf	$\sim 3min$	$60epochs$
SpecLSTM	$\sim 1h30$	$52epochs$
SpecBLSTM	$\sim 3h45$	$52epochs$
TS-LSTM	$\sim 7h$	$52epochs$
TS-BLSTM	$\sim 15h$	$52epochs$

BLSTM networks give better results. Compared to the E2-SVMrbf computation time per fold (~3min), they are still slow : in E2-SVMrbf, each utterance is summarized with a vector of statistical functions whereas for other models, input data are large bidimensional arrays (spectrograms and time series).

We can also observe BLSTM computation has nearly doubled in time compared to LSTM computation and has nearly four times higher with time series input data than with spectrograms input data.

Conclusion

We proposed to compare SVM vs LSTM neural networks trainings using different features methods: a minimalistic E2-334-LIMSI set of voice parameters, time series of these voice parameters, and spectrograms. We obtained quite similar score results with different approaches. Some improvements in a future work could be performed such as attention mechanism and CTC loss function.

References

- Eyben, and al. 2016 The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing.
- Chernykh and al. 2017. Emotion recognition from speech with recurrent neural networks.
- Ghosh and al. 2015. Learning Representations of Affect from Speech.
- Han, and al. 2014. Speech emotion recognition using deep neural network and extreme learning machine.
- Lee and Tashev. 2015. High-level feature representation using recurrent neural network for speech emotion recognition.
- Schuller et al. 2017. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals.
- Busso et al. 2008. Iemocap: interactive emotional dyadic motion capture database.