Supplementary material

1 Description of ComParE2016 features

The ComParE2016 [1] features are computed using a frame rate of 100 Hz. All features are computed using a window size of 20 ms, except the zero crossing rate, which uses a 60 ms window size. A hamming window is used for the spectrogram computations.

In the discussion below, we use the following notations: a discrete time signal is denoted by $x[t]$ with t as the time index, a time-frequency representation is denoted by $X[n, k]$ with n denoting the frame index and k denoting the frequency bin index. The index n is omitted, wherever not necessary. The FFT size is denoted by K and frequency at k^{th} bin

$$f[k] = k/K \times \frac{\text{sampling rate}}{2} \tag{1}$$

1.1 Root-mean-square (RMS) energy

Let, $x[t]$ be the sequence of discrete time samples in a frame of N samples. The RMS energy is computed as,

$$\text{RMS Energy} = \sqrt{\frac{1}{N}\sum_{t=0}^{N-1} x[t]^2} \tag{2}$$

1.2 Zero crossing rate (ZCR)

ZCR is defined as the average number of times the signal changes its polarity, that is the signal value changes from positive to negative or vice-versa, in a second. ZCR is calculated after correcting for the mean signal drift from zero.

1.3 Spectrogram

A spectrogram is a time-frequency representation, computed as the discrete Fourier transform (DFT) of the windowed signal at regular intervals. Let $w[t]$, $[0 \leq t \leq T-1]$, denote a T length window. The spectrogram $X[n, k]$

of signal $x[n]$ is computed as,

$$X[n,k] = \frac{1}{T}\sum_{t=0}^{T-1} w[t]x[t+nT_s]\exp\left(-j2\pi kt/T\right), \quad 0 \leq k \leq K-1,\ 0 \leq n \leq T_x/T_s \quad (3)$$

where T_s is the window shift, K is the DFT size, and T_x is the number of samples in $x[t]$.

1.4 Mel spectrogram

Mel spectrogram is the power spectrogram representation ($|X[n,k]|^2$) of the signal mapped to the Mel frequency scale. Squared magnitude of the spectrogram is weighted using the Mel filter weights and added to obtain the Mel spectrogram representation. In the ComParE2016 feature set, the Mel spectrogram is computed using 26 frequency bands in the frequency range 20 Hz-8 KHz.

1.5 Auditory spectrogram

The auditory spectrogram is obtained by applying auditory pre-emphasis and loudness compression to the Mel spectrogram. Auditory pre-emphasis is done by multiplying the Mel spectrogram with the equal loudness curve and the loudness compression is done according to the power law of hearing (cubic root). Logarithm of the auditory spectrogram is used as the feature in the ComParE2016 feature set.

1.6 Modulation-filtered auditory spectrogram

RASTA filter [5] is the used as the Modulation filter and applied to the log-auditory spectrogram.

1.7 Mel frequency cepstral coefficients (MFCCs)

MFCCs [6] are computed by applying the discrete Cosine transform (DCT) to the logarithm of the Mel-spectrogram of the signal. In the ComParE2016 feature set, the first 14 coefficients of the DCT output are taken as the MFCC features.

1.8 Spectral roll-off point

Spectral roll-off point is defined as the frequency below which a pre-defined percentage p of the spectral energy is concentrated, i.e.,

$$\text{Spectral roll-off point} = \arg\max_i \left[\frac{\sum_{k=0}^{i}|X[k]|^2}{\sum_{k=0}^{K-1}|X[k]|^2} \leq p\right] \quad (4)$$

In the feature set, 4 spectral roll-off points are computed at $p \in [0.25, 0.5, 0.75, 0.9]$.

1.9 Spectral centroid

Spectral centroid is computed as the weighted mean of the frequencies with the power of the spectral components as weights,

$$\text{Spectral centroid } (\mu) = \frac{\sum_{k=0}^{K-1} f[k] \left|X[k]\right|^2}{\sum_{k=0}^{K-1} \left|X[k]\right|^2}. \tag{5}$$

1.10 Spectral variance

Spectral variance is the second moment of the spectrum, and computed as

$$\text{Spectral variance } (\sigma^2) = \sqrt{\frac{\sum_{k=0}^{K-1} \left(f[k]-\mu\right)^2 \left|X[k]\right|^2}{\sum_{k=0}^{K-1} \left|X[k]\right|^2}}. \tag{6}$$

1.11 Spectral skewness

Skewness is the third moment of the spectrun and computed using

$$\text{Spectral Skewness} = \sqrt{\frac{\sum_{k=0}^{K-1} \left(f[k]-\mu\right)^3 \left|X[k]\right|^2}{\sigma^3 \sum_{k=0}^{K-1} \left|X[k]\right|^2}}. \tag{7}$$

1.12 Spectral kurtosis

Kurtosis is the fourth moment of the spectrum and computed as

$$\text{Spectral Kurtosis} = \frac{\sum_{k=0}^{K-1} \left(f[k]-\mu\right)^4 \left|X[k]\right|^2}{\sigma^4 \sum_{k=0}^{K-1} \left|X[k]\right|^2}. \tag{8}$$

1.13 Spectral slope

Spectral slope is computed as

$$\text{Spectral slope} = \frac{\sum_{k=0}^{K-1} \left(f[k]-\mu_f\right) \left(\left|X[k]\right|^2 - \mu\right)}{\sum_{k=0}^{K-1} \left(f[k]-\mu_f\right)^2}. \tag{9}$$

1.14 Spectral flux

Spectral flux is computed as the RMS value of the magnitude spectrum differences between successive time frames,

$$\text{Spectral flux} = \sqrt{\frac{1}{K} \sum_{k=0}^{K-1} \left| \left|X[n,k]\right| - \left|X[n-1,k]\right| \right|^2}. \tag{10}$$

1.15 Spectral entropy

Spectral entropy is defined as the entropy of the normalized power spectrum. The discrete frequency power spectrum is first mapped to a probability mass function, followed by entropy computation. Spectral entropy is computed as,

$$\text{Spectral entropy} = -\sum_{k=0}^{N-1} \hat{X}[k] \log_2 \hat{X}[k], \tag{11}$$

where $\hat{X}[k] = \frac{|X[k]|^2}{\sum_{k=0}^{N-1} |X[k]|^2}$.

1.16 Spectral energy in bands

Spectral energy in a frequency band with lower and upper cut-off frequencies f_1, f_2 respectively is computed as,

$$\text{Spectal energy} = \sum_{k: f_1 <= f[k] <= f_2} |X[k]|^2. \tag{12}$$

Spectral energies in bands $250 - 650$ Hz and $1 - 4$ KHz are computed in the ComParE2016 feature set.

1.17 Psychoacoustic sharpness

Psychoacoustic sharpness is defined as the spectral centroid computed on the bark scale. The magnitude spectrogram is first mapped to the Bark frequency scale and the centroid is computed using eqn. 5.

1.18 Harmonicity

Harmonicity is computed as the ratio of the energy at the peaks in the spectrum to the total energy.

1.19 F0

F0 refers to the fundamental frequency of the signal. The sub-harmonic summation (SHS) method [4] is used to compute F0, which is further refined using Viterbi smoothing and voicing detection.

1.20 Probability of voicing

Probability of voicing measures the probability of the signal segment being voiced. A value close to zero indicates un-voiced signal and a value close to one indicates a perfectly periodic signal. It is computed as,

$$v = \max\left(0, 1 - \frac{\text{Average energy of F0 candidates}}{\text{energy at the chosen F0}}\right). \tag{13}$$

1.21 Log. Harmonics-to-Noise Ratio (HNR)

HNR is defined as the ratio of the energies of periodic to non-periodic components in the signal. HNR is computed using the auto-correlation function (ACF) in the ComParE2016 feature set. The value of ACF at the lag corresponding to F0 is taken as the energy of the periodic component and the value at lag 0 is the sum of periodic and non-periodic component energies.

1.22 Jitter

Jitter is defined as the average absolute difference of pitch between successive time frames. Two values, jitter and its delta over time are computed in the ComParE feature set.

1.23 Shimmer

Shimmer is defined as the average absolute difference of peak amplitude between the consecutive periods divided by the average amplitude.

2 Statistics computed on the LLDs

The list of statistics, referred to as functionals, computed on the low-level feature descriptors is given in Tab. 1. A set of 100 different statistics are computed for the spectral and energy related LLDs given in Tab. 2. For voicing related LLDs, except the F0 feature, a total of 78 statistics are computed and for the F0 feature 83 functionals are computed.

The set of functionals computed per LLD, and the list of LLDs is available in the configuration files[1][2] used by the opensmile toolbox.

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 Interspeech, 2016, pp. 2001–2005.[Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-129.

[2] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM International Conference on Multimedia, ser. MM '10.New York, NY, USA: Association for Computing Machinery, 2010,p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246.

[1] github.com/naxingyu/opensmile/blob/master/config/ComParE_2016_core.lld.conf.inc
[2] github.com/naxingyu/opensmile/blob/master/config/ComParE_2016_core.func.conf.inc

[3] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," Frontiers in Psychology, vol. 4, p. 292, 2013. [Online].Available: https://www.frontiersin.org/article/10.3389/fpsyg.2013.00292

[4] Hermes, Dik. (1988). Measurement of pitch by subharmonic summation. The Journal of the Acoustical Society of America. 83. 257-64. 10.1121/1.396427.

[5] Hermansky, H., and Morgan, N. (1994). RASTA Processing of Speech. IEEE Transactions on Speech, and Audio Processing. Vol 2 (4).

[6] Davis, S. B., and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pp. 357–366.

Table 1: List of statistical features derived from low-level descriptors given in Table 2 [3].

Statistic features derived from LLD	No. Func.	Spectral & Energy		Voicing		Group
		LLD	Δ LLD	LLD	Δ LLD	
Quartiles 1–3, 3 inter-quartile ranges	6	✓	✓	✓	✓	Percentiles (9)
1% percentile (∼ min), 99% percentile (∼ max)	2	✓	✓	✓	✓	
Percentile range 1 %–99 %	1	✓	✓	✓	✓	
Position of min / max, range (max – min)	3	✓	✓	✓	✓	Temporal (15)
Contour centroid	1	✓		✓	✓	
Contour flatness	1	✓	✓	✓	✓	
Rel. duration LLD is > 25 / 50 /75 /90% range	4	✓	✓	✓	✓	
Relative duration LLD is rising	1	✓	✓	✓	✓	
Relative duration LLD has positive curvature	1	✓	✓	✓	✓	
Mean, max, min, std. deviation of segment length	4	✓	✓*			
Mean value of peaks	1	✓	✓			Peaks (12)
Mean value of peaks – arithmetic mean	1	✓	✓			
Mean / std.deviation of inter peak distances	2	✓	✓			
Amplitude mean of peaks, of minima	2	✓	✓			
Amplitude range of peaks	2	✓	✓			
Mean / std. deviation of rising / falling slopes	4	✓	✓			
Arithmetic mean	1	✓		✓	✓	Moments (6)
Root quadratic mean	1	✓	✓	✓	✓	
Positive arithmetic mean	1		✓	✓	✓	
Standard deviation, skewness, kurtosis	3	✓	✓	✓	✓	
Linear regression slope, offset, quadratic error	3	✓		✓	✓	Regression (7)
Quadratic regression a, b, offset, quadratic error	4	✓		✓	✓	
Linear prediction (LP) gain, LP Coefficients 1–5	6	✓	✓	✓	✓	Modulation (6)

1. For F0 LLD, additionally, the 4 segment length related LLDs and 1 percentage of non-zero frames functionals are applied.
2. Total number of functionals applied is 100 for spectral and energy LLDs, 78 for voicing related LLDs except F0, and 83 for F0.

Table 2: The set of low-level descriptors computed in the ComParE2016 feature set [2].

	Low level descriptor (LLD)	Dim.
Energy	RMS Energy, Zero-Crossing Rate	2
	Sum of modulation-filtered auditory spectrum	1
	Sum of auditory spectrum (loudness)	1
Spectral	modulation-filtered auditory spectrogram (0–8 kHz)	26
	Mel frequency Cepstral coefficients (MFCCs)	14
	Spectral Flux, Centroid, Entropy, Slope	4
	Psychoacoustic Sharpness, Harmonicity	2
	Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	4
	Spectral Variance, Skewness, Kurtosis	3
	Spectral energy 250–650 Hz, 1 kHz–4 kHz	2
Voicing	F0 (SHS & Viterbi smoothing)	1
	Log. HNR, Jitter (local, Δ), Shimmer (local)	4
	Probability of voicing	1