WAVELET METHOD FOR BREATH DETECTION IN AUDIO SIGNALS

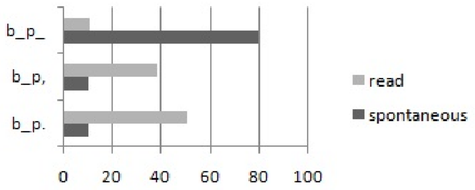
Magdalena Igras, Bartosz Ziółko
AGH University of Science and Technology, Cracow, Poland

Abstract

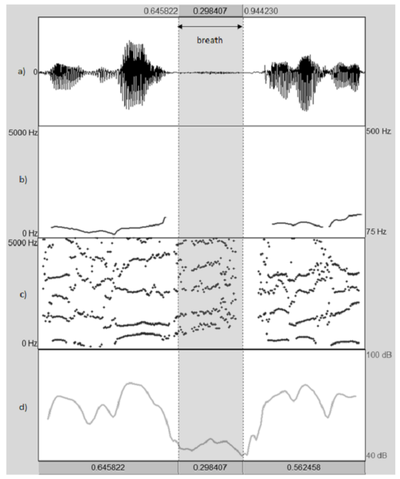
The paper presented a method of automatic breath detection in audio. The algorithm uses both temporal and spectral features. It is based on method of wavelet decomposition with perceptual scale, used in ASR dedicated for Polish. Evaluation tests of the algorithm and human perception of breath in audio were performed nad discussed. The research has multiple application in both speech technology (improving performance of ASR, supporting punctuation detection in speech, improving recordings quality) and potentially in medicine (monitoring breath disorders or supporting detection of emotional arousal).

Background

❑ In normal physiological condition, at rest, the value of breath per minute is 12-20, while during speaking (dynamic respiratory), it is 10-12.
❑ Breaths in audio can be considered as a disadvantage that interfere clarity of the speech signal, or on the other hand - as an integral part of the signal, that carries information on:
• punctuation,
• oratorical abilities of the speaker,
• emotions
• physical condition of the speaker.



Contribution of breath pauses to determining full stops and commas - comparison between read and sponta-neous speech; b_p_ - percent of breaths not correlated withfullstops nor commas, b_p. - percent of breath pauses denoting fullstops, b_p, - percent of breath pauses denoting commas



A part of a speech signal containing a breath: a) wave-form, b) F0 contour, c) formants d) energy contour

1 Data

❑ monologues in formal or half-formal situations, both read and spontaneous speech.

❑ experienced speakers (politicians, proffessors, translators) and unexperienced speakers (students) 24 speakers (13 male, 11 female)

❑ 30 min of Polish utterances from orations or public presentations: speeches and reports from Europarlament , sessions of faculty council, students lectures and reviews + 30 min of recordings of real time translation of orations during Europarlament sessions.

Total duration of recordings is 60 min,
1 to 5 min for each speaker.

Training set: 30 min
Testing set: 30 min

2 Temporal features

The threshold of the cryterias are constituted on the base of statistical analysis of the training set of recordings:

❑ durations

# breaths	mean [s]	min [s]	max [s]	sd [s]
363	0.446	0.101	2.478	0.368

Statistical parameters of breaths length in audio

❑ energy

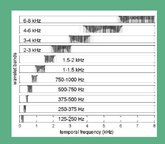
# breaths	mean	min	max	sd
363	1.243	0.142	14.168	1.584

Statistical parameters of breaths energy in audio

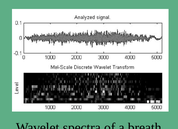
❑ lack of fundamental frequency

3 Wavelet features

For each of the preliminary hypotheses of breaths, discrete wavelet transform is computed. The signal is decomposed in 12 resolution levels using perceptual scale.



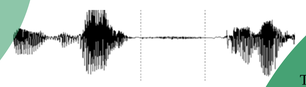
Width of frequency subbands in speech spectra



Wavelet spectra of a breath

Preliminary hypotheses of breath regions

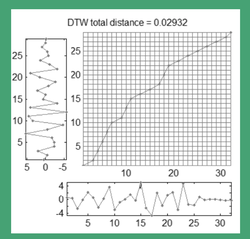
Breath detection algorithm



4 Classification

To establish similarity of the hypothesis to the breath template, we use dynamic time warping algorithm (DTW).

It allows to obtain the distance measure between the two signals of different lengths.



An example DTW path between two vectors representing energy parameters in one decomposition level

Final hypotheses of breath regions

5 Breath hypotheses

For each tested recording, the temporal and wavelet parametrisation was performed.

The result was the features matrix for each hypothesis, that consisted of 12 parameters for each frame of the signal.

Then, for the parameters for each decomposition level, DTW was computed, normalized and the results were summed.

The mean value of distances of the tested signal and all the patterns determinated if the the hypothesis was accepted or rejected.

The final results were compared and we counted number of:
- true positives,
- false positives,
- false negatives

in reference to the annotation.

Results

❑ Results of algorithm evaluation:

# breaths	tp	fn	fp	R	P	F
340	340	19	513	40.0	94.7	56.1

❑ Good precision, poor recall => is the algorithm too sensitive?
❑ To verify this, we perfomed perceptual tests of isolated breaths, detected by the algorithm.
❑ The results of perceptual tests induce also that human perception easily detects long breaths between phrases in audio, but skips the short ones. Listeners recognize them only when isolated from the speech in recordings.
❑ Most of false positives were short breaths between words, while correct results were typical long breaths separating the phrases of the utterance.

Literature

E. Shriberg, A. Stolcke, D. Hakkani-T ¨ur, and G. T¨ur, “Prosody-based automatic segmentation of speech into sentences and topics,” 2000.

Ikuo Homma and Yuri Masaoka, “Breathing rhythms and emotions.,” Experimental physiology, vol. 93, no. 9, pp. 1011–1021, Sept. 2008.

Dima Ruinskiy and Y. Lavner, “An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 3, pp. 838–850, 2007.

Tomoyasu Nakano, Jun Ogata, Masataka Goto, and Yuzuru Hiraga, “Analysis and automatic detection of breath sounds in unaccompanied singing voice,” 2008, pp. 387–390, ICMPC 2008.

M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, and J. Wicijowski, “Automatic speech recognition system based on wavelet analysis,” In 2010 IEEEFourth International Conference on Semantic Comput-ing, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, pp. 450–451, 2010.

Applications

- improving quality of recordings by removing breath,
- audio information retrieval,
- inserting punctuation in automatic transcripts,
- support for detection of emotion or physical effort level,
- diagnostics of potential respiratory
- disfunctions (e.g.sleep apnea),
- measure of physical fitness,
- measuring speech quality.

The research was funded by the National Science Centre allocated on the basis of a decision DEC-2011/03/D/ST6/00914