Shout analysis and characterisation

Larbi Mesbahi, David Sodoyer, Sébastien Ambellouis

Shout analysis and characterisation

Larbi Mesbahi[1] · David Sodoyer[1] · Sebastien Ambellouis[1]
[1] LEOST, IFSTTAR , Lille, France

February 7, 2019

Abstract

The goal of this paper is to analyse and to characterise the shout of a people to develop an automatic detector. We define a shout as a voiced part of an audio signal maintained over time. We show that a set of formants parameters can be defined to discriminate a typical "shout" from a "neutral" part of a speech. Moreover, it appears clearly that the duration of the window used to estimate these parameters is critical to yield better results. We conclude by presenting a performance analysis in the noisy context of a transport surveillance application.

Index Terms: Shout/Speech discrimination, formants analysis

1 Introduction

The systems of detection and classification of sound events are considered as innovative applications in intelligent system domain [1], in man-machine interaction [2], in transport [3],..etc. The sound event studied in this paper concerns the shout of a people. In this case, the objective of such device is often the release of automatic alarms to give a help to persons in difficulty.

This topic has interested many researchers: we can cite [4], the author was interested to the shout detection among other sound events by using cepstral parameters. In [5], the authors mention the importance to differentiate between a shout and a shouted speech. Several works as [6], [7], [8] has developed systems of shout detection based on MFCC parameters and F_0 using HMM and SVM models.

The approach proposed by [9] can be used in preliminary phase to reinforce the detection of speech or shout. So, the Line Spectral Frequency (LSF) parameters are strong to be used in voice activity detection, mainly when we are constrained to noise. In second phase, a shout/speech detection system can be applied. In [8], the authors has described the shout by the parameter of continuity of energy. In papers [10] and [11], the shouted speech is characterised in particular by open and close glottal proprieties. The authors has noted that a shouted speech can be differentiated by their low variance of energy in low frequencies.

Several studies such that [12], [10] and [11] defined the speech into five levels: whishpered, soft, neutral, loud and shouted. By considering the last level, [5]

recall that is important to differentiate between a shouted speech and a shout taking into account the existence of a linguistic context or not. Finally, the shout which can be realized in various contexts (physical suffering, aggression, calling for help, sadness, surprise, enjoyment, signs of presence, ..etc.), it is often associated with an emotion. The work of [13] is an illustration of the automatic classification of real-life emotions.

In this paper we are interested to analyse and characterise the shout without considering the linguistic context nor the associated emotion. A shout is defined as a voiced part of a signal with a few articulation and a strong energy maintained in time. In first step, we aim to describe the shout in a space of parameters allowing to differenciate between a shout and a speech. The discrimination to others sound events are not considered in this paper. However, the noise is taken into account to test the reliability of our parameters in realistic noised situations. This paper is organized as follows. The section 2 presents the parameters of discrimination between a shout and a speech, then the analysis steps which follow. The section 3 describes the corpus and the configuration of studied parameters. The results are presented in section 4. Finally, the section 5 is dedicated to a conclusion and some perspectives.

2 Analysis and characterisation of shout

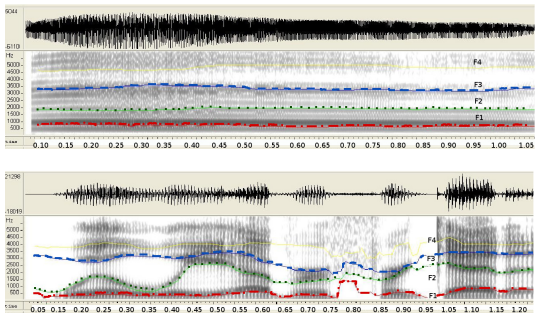


Figure 1: Presentation of the shout signal (top) and the speech signal (down) and theirs spectrograms including the different formants.

We define a shout as a voiced part of a speech signal and doesn't contain a linguistic information. We perform a comparative formant analysis between shout and speech which allows to characterize the principal parameters of the spectral envelope of the voiced sounds. The figure 1 shows the spectrogram and the first four formants of the shout and speech signals. We can observe that the first three formants of the shout are stable in frequency and energy relative to the speech. The variability of speech formants is justified by the articulation of vowels and consonants and the sequence of voiced parts. In context of the shout as we have defined, these co-articulations are relatively

weak nonexistent. Studying the variance (or standard deviation) in frequency or in energy for each extracted formant seems a pertinent solution to discriminate the segments of shout and speech. The variances must be defined on a duration sufficient to take into account the different kind of shout. For that purpose, we suggest analyzing the variance (standard deviation more exactly) of the formant frequencies and that of the energy carried by each of them according to the duration of observation. In this way, for a formant F_i and a duration of observation N (expressed in number of observation frames), the standard deviation $\sigma_{f_{Fi}}$ is defined by the following equation:

$$\sigma_{f_{Fi}} = \sqrt{\frac{1}{N}\sum_{n=0}^{N}[f_{Fi}(n) - \mu_{f_{Fi}}]^2} \quad and \quad \mu_{f_{Fi}} = \frac{1}{N}\sum_{n=0}^{N} f_{Fi}(n) \tag{1}$$

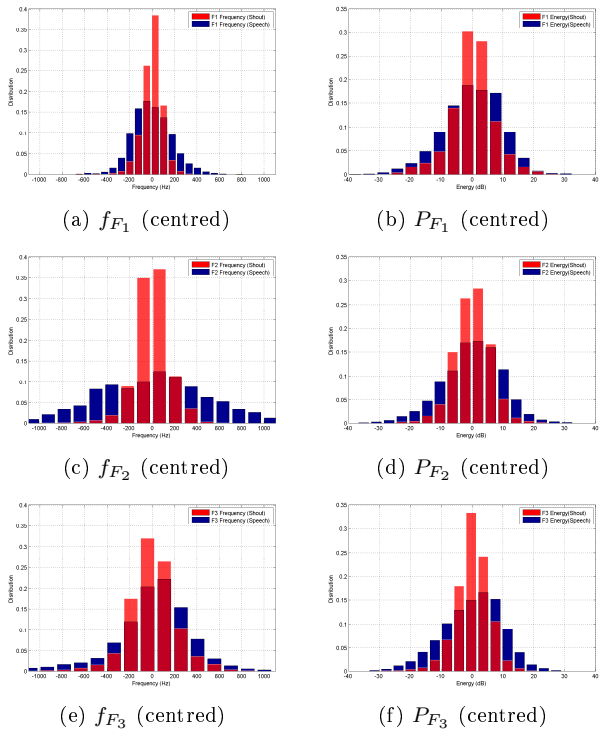
and the $f_{Fi}(n)$ is the F_i formant frequency for the n^{th} frame. The standard deviation $\sigma_{P_{Fi}}$ of the power spectral density P_{Fi} for the F_i formant is defined as following:

$$\sigma_{P_{Fi}} = \sqrt{\frac{1}{N}\sum_{n=0}^{N}[P_{Fi}(n) - \mu_{P_{Fi}}]^2} \quad and \quad \mu_{P_{Fi}} = \frac{1}{N}\sum_{n=0}^{N} P_{Fi}(n) \tag{2}$$

3 Corpus and methodology

The aim of this study is to analyze the first, the second and the third formant, respectively F1, F2 and F3. Our objective is to make a decision function in order to distinguish between a shout event and a speech event. For this purpose a database containing audio recordings of shout and speech are necessary. For mention, our database was built in our laboratory. So, it was difficult to acquire a corpus containing different variants of shout signal and not altered by noise. For this purpose, we have collected a database containing differents audio recordings of shout and speech. First part, we collected from different web sources 91 examples of shout recordings. These shout examples can take different forms : screaming, panic, baby cry, pain, etc. Each example have a duration between 0.11 seconds and 5 seconds. The mean duration of examples is 0.7 seconds. After verification, we have retained only the samples that respect a good quality. In second part, the speech examples are extracted from the the CMU Arctic database [14], we have selected 27 examples of male and female voices. The duration of each speech example varies between 1.09 seconds and 6.42 seconds. The speech examples contain voices of male and female equitably. The texts are pronounced in English. The mean duration of examples is 3.5 seconds.

After collecting the data , we proceed to formant extraction. So the formant frequencies (F_1,F_2,F_3) are estimated with the Wavesurfer [15]. First we have perform a pre-emphasis operation, then the algorithm estimates the LPC parameters each 10 ms using a Hamming window of 50 ms. We mention that the



(a) f_{F_1} (centred) (b) P_{F_1} (centred)

(c) f_{F_2} (centred) (d) P_{F_2} (centred)

(e) f_{F_3} (centred) (f) P_{F_3} (centred)

Figure 2: Histograms centered and normalized of the formants frequencies (F_1 F_2 F_3) (1^{st}column) and their powers (2^{nd}column) corresponding to the shout and the speech

dynamic programming allows to estimate the formant frequencies based on the LPC coefficients. After this operation, a manual correction of boundaries was performed to correct errors of estimation. The powers of formants were calculated with a power spectral analysis based on the LPC parameters. Finally, we obtained all the F_1,F_2 and F_3 formant frames for shout and speech.

4 Results and discussions

4.1 Standard deviation of formants in frequency and in energy

Figures 2 presents respectively centered and normalized histograms of frequencies and powers taken for the first three formants of the corpus (shout and speech). The centered character was obtained by removing the mean of each of parameters fF_i and PF_i, estimated on the whole corpus. The observation

# frames N	10 (0.1s)	30 (0.3s)	50 (0.5s)	70 (0.7s)	100 (1s)
# shout frames $\geq N.10ms$	100%	80%	46%	30%	21%

Table 1: Percentage of shout signals with duration $\geq$ to the duration of observation N.10ms

made on all the spectrograms of the figure 1 is generalized on all examples of the corpus. For each cases (frequency or power), the histograms associated in shouts present standard deviations more lower than those associated to the speech data. The next section present the impact of durations of the observation on standard deviations of frequencies and powers of the first three formants.

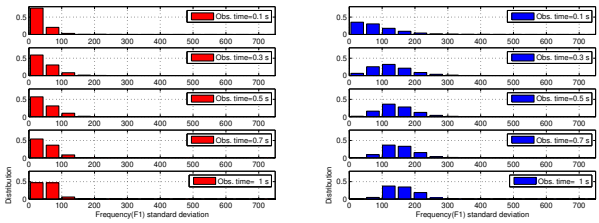
4.2 Duration of the observation

The different variances were estimated on the whole corpus (shout/speech) on durations of observation of $N = 10$, 30, 50, 70 and 100 consecutive frames. Formants being estimated each 10 ms, these durations of observations correspond respectively to times of observations of 100, 300, 500, 700 ms and 1 second. Some shouts being sometimes shorter than the duration of observations and to avoid the bias of the estimation, these shouts were discarded. The impact is a reduction of the number of shout signals used in the estimation.

The table 1 indicates the percentage of retained shouts signals according to the duration of observation.

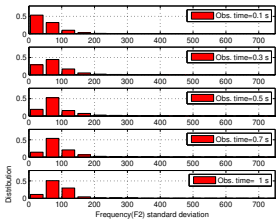
Figure 3 presents the distribution of the standard deviation of frequencies according to the duration of observation, for respectively the shout and the speech. We are interested first to the standard deviation relative to speech. We can observe that the standard deviations of F_1 are much lower than those of the other formants. This is coherent because in a general way the F_1 formant has a dynamic frequency lower than F_2 and F_3. The most low standard deviations correspond to the small observation window (100 ms); this is can be explained by the fact that the observation window has a duration lower than the average duration of a vowel or a syllable CV/VC. From 300 ms the mean of standard deviations increase until 500 ms where it stabilizes; the duration of observation corresponds to the duration of several consecutive syllables. As regards the shout, the evolution of histograms with respect to N is slightly different. At 100 ms, whatever the formant, the difference between speech and shout is certainly due to a very low representation of the syllables CV/VC, limiting itself in great majority to the vowels (histograms are more compact than those of the speech at 100 ms). Then, the syllabic content remaining close to a vowel and histograms of standard deviations remain compact with averages stabilizing from 500 ms. Generally speaking, for the same formant, the distribution of standard deviations of the shout and the speech recover slightly for 300 ms. This recovery disappears from 500 ms, allowing to discriminate better between the shout and speech whatever the considered formant.

These observations are the same on the distribution of standard deviations

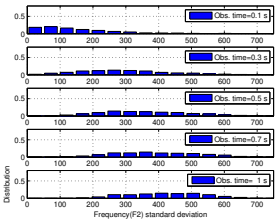


(a) f_{F_1} (Shout)

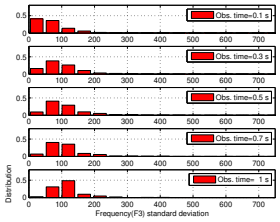
(b) f_{F_1} (Speech)



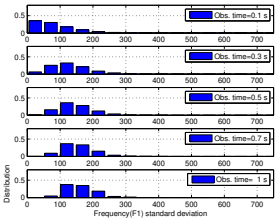
(c) f_{F_2} (Shout)



(d) f_{F_2} (Speech)



(e) f_{F_3} (Shout)



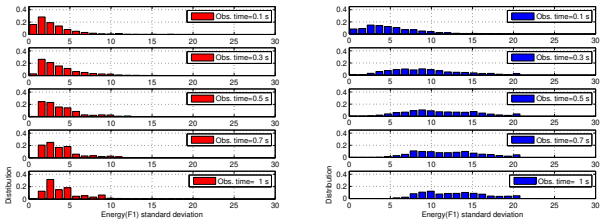
(f) f_{F_3} (Speech)

Figure 3: Distributions of standard deviation of formant frequencies (F_1 F_2 F_3) corresponding to the shout (1^{st}column) and the speech (2^{nd}column)

of powers for each of the 3 formants for the shout (figure 4 1^{st}column) and for the speech (figure 4 2^{nd}column)).

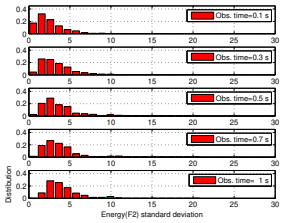
4.3 Detection in neutral and noised mode

The analysis presented previously shows that it is possible to differenciate between a shout and a speech by analyzing the distribution of the standard deviation in frequency and in power for the first three formants extracted from the signal. It seems that is possible to conceive a detector which allows to segment a sound signal in two classes "shout" (sh) and "speech" (sp). A class is assigned for each frame of the signal after comparing the standard deviation of the frequency f and the energy P with a given threshold θ. A frame n is considered

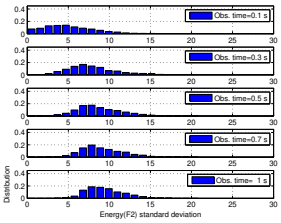


(a) P_{F_1} (Shout)

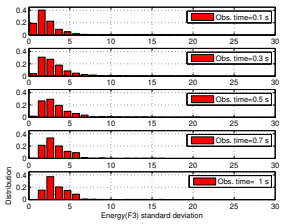
(b) P_{F_1} (Speech)



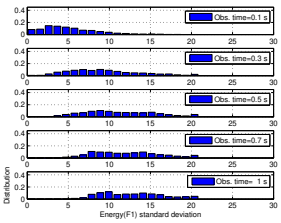
(c) P_{F_2} (Shout)



(d) P_{F_2} (Speech)



(e) P_{F_3} (Shout)



(f) P_{F_3} (Speech)

Figure 4: Distributions of standard deviation of powers of formant frequencies (F_1 F_2 F_3) corresponding to the shout (1^{st}column) and the speech (2^{nd}column)

as a shout ($\mathcal{H}1$ hypothesis) if

$$\sigma_{x_{F_i}} \underset{\mathcal{H}0}{\overset{\mathcal{H}1}{\lesseqgtr}} \theta_{x_{F_i}} \tag{3}$$

Where $\sigma_{x_{F_i}}$ is the standard deviation of the analysed frame n, θ is the standard deviation threshold, $x \in f, P$ and $i \in [1, 2, 3]$.

In this case, formants is extracted automatically without a manual correction. For each kind of parameters (standard deviation in frequency f and in energy P) and for each formant, we have varied the $\sigma_{x_{F_i}}$ value in object to establish the ROC curves representing the rate of good detection according to the rate of false detection. The rate of false detection is defined as the ratio of the number of speech frames detected as shout to the total number of speech frames. The rate of good detection is defined as the ratio of the number of shout frames detected as shout to the total number of shout frames. The ROC

curve is established on several durations of observation N. It is necessary to note that the total number of frames to be detected evolves according to the N value because some shout examples have a duration inferior than the duration of observation. For each duration the maximal number of detectable shouts of the corpus is precised in the table 1.

The results are presented in figure 5 for the parameters $\sigma_{f_{F_i}}$ (in first column) and $\sigma_{P_{F_i}}$ (in second column). By considering the figure 5 on the first column, the results are encouraging for F_1 and F_2. In spite of some estimation errors added naturally to variations of $\sigma_{f_{F_i}}$, we obtain suitable rates of good and bad detection of F_1 and F_2 from duration of 0.5s ($N = 50$). These rates increase/decrease with the increase of N. Concerning the $\sigma_{f_{F_3}}$ parameter the performances decrease because the estimation of F_3 is strongly noised. With regard to figure 5 (second column), we observe the same behaviour of the detector. The errors of estimation generated on F_1, F_2 and F_3 affect an unimportant part of detector performances. At 500ms we obtain scores of BD/FA around 0.9/0.07 for F_1 and 0.85/0.1 for F_2. For F_3, although the standard deviation increases in frequency (see figure 5), the impact on the associated power values is low and the performances are equivalent to the first two formants.

Finally, the best performances are obtained for a duration of 1s. Unfortunately, this choice of duration obliges us to reject 79% of shout examples. At 500ms, we hope detecting a score around 46% of total shouts. This seems low but the evaluation is done frame by frame. The challenge in the implementation of the detector is to choose a formant F_i, a parameter ($\sigma_{f_{F_i}}$ or $\sigma_{P_{F_i}}$) and a threshold $\theta_{x_{F_i}}$ $(x \in f, P)$, giving as result a score of 0.8/0.01 (e.g. $\sigma_{f_{F_2}}$) with a duration allowing to detect more possible shouts (e.g. $N = 30$). Thus, we can interpret generally that we are able to detect correctly 80% of shout frames representing 80% of total shout examples.

5 Conclusions and future work

We have showed that is possible to characterise the signals of shout and speech by analyzing the temporal evolution of parameters at low level such the formants. The results are encouraging and can be improved in term of detection shout/speech. This can be realized by merging different temporal evolutions, in a new model or at the decision step. It remains to study the performances of such detector in presence of other sound events different to speech and shout. In future work, if needed, other parameters of high level can be diverted from this study and it is possible to inject apriori knowledges for a complex modelling (e.g. HMM "Hidden Markov Model", SVM "Support Vector Machine" and GMM "Gaussian Mixture Model"). In perspective to improve the actual results, a logistic function can be used as a binary decision function. On the other hand, in order to evaluate the efficiency of our detection model, the technique of neural networks can be applied and eventually compare the different results.

References

[1] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J. Seri, "Sound and speech detection and classification in a health smart home," in *Engeneering in Medicine and Biology society. EMBS 200. 30 the Annual International Conference and Applications*, vol. 2, 2008, pp. 4644–4647.

[2] M. Janvier, X. Alameida-Pineda, L. Girin, and R. Horaud, "Sound event recognition with a compagnion humanoid," in *IEEE International Conference on Humanoid Robotics, Humanoids*, Osaka, Japan, 2012.

[3] Q.-C. Pham, A. Lapeyronnie, C. Baudry, L. Lucat, P. Sayd, S. Ambellouis, D. Sodoyer, A. Flanquart, A.-C. Barcelo, E. Heer, F. Ganansia, and V. Delcourt, "Audio video surveillance system for public transportation," in *2nd International Conference on Image Processing Theory Tools and Applications (IPTA)*, Paris, France, 2010, pp. 47–53.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP Journal*, vol. 2009, 2009.

[5] H. Nanjo, T. Nishiura, and H. Kawano, "Acoustic-based security system:towards robust understanding of emergency shout," in *Proc. Int. Conf. Inf. Assurance and Sec.*, 2009, pp. 725–728.

[6] L. D. Besacier, A. A., and P. F. M., "Automatic sound recognition relying on statistical methods, with application to telesurveillance," in *Proceedings of COST 254, International Workshop on Intelligent Communication Technologies and Applications, with Emphasis on Mobile Communication*, May 5-7, 1999.

[7] D. Liderman, A. Cohen, E. Zmora, K. W. rmke, S. Hauschildt, and A. Stellzig-Eisenhauer., "Automatic classification of the cry of infants with cleft palate," in *2nd European Medical and Biomedical Engineering Conference in Vienna*, 2002.

[8] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *IEEE conference on Industrial Electronics and Applications*, vol. 2, 2010, pp. 2115–2120.

[9] H. M. M. O. C. S. P. Roy, "Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal," *International Journal of Speech Technology*, vol. 21(4), pp. 753–760, 2018.

[10] V. K. Mittal and B. Yegnarayana, "Effect of glottal dynamics in the production of shouted speech," *Acoustical Society of America journal*, vol. 133(5), pp. 3050–3061, 2013.

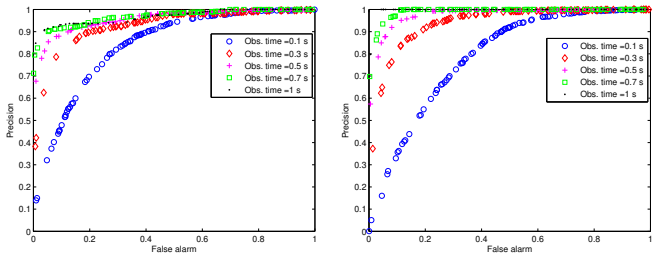
[11] V. K.Mittal and B. Yegnarayana, "Production features for detection of shouted speech," in *Consumer Communications and Networking Conference (CCNC), IEEE*, 2013, pp. 106–111.

[12] C. Zhang, "Analysis and classification of speech mode: Whispered through shouted," in *INTERSPEECH 2007*, Antwerp, Belgium, 2007, pp. 2289–2292.

[13] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *IEEE-Trans. Information Theory*, vol. 50(6), pp. 487–503, 2008.

[14] J. Kominek and A. Black, *The CMU ARCTIC speech databases for speech synthesis research*, 2003.

[15] in *http://www.speech.kth.se/wavesurfer/*.

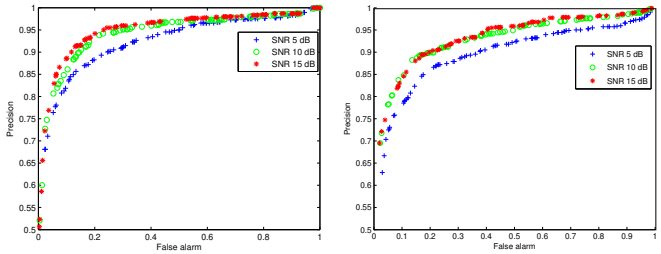


(a) $\sigma_{f_{F_1}}$ (b) $\sigma_{P_{F_1}}$

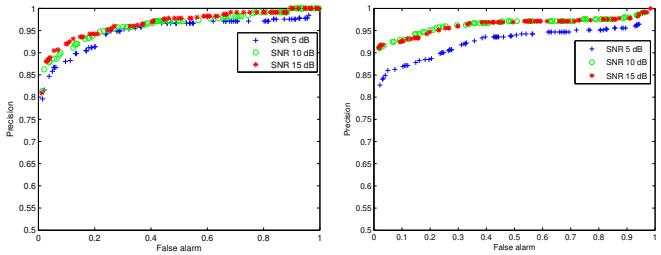
(c) $\sigma_{f_{F_2}}$ (d) $\sigma_{P_{F_2}}$

(e) $\sigma_{f_{F_3}}$ (f) $\sigma_{P_{F_3}}$

Figure 5: ROC curves based on frequency and power for the detection of shout vs speech in neutral mode.



(a) $\sigma_{f_{F_2}} N = 30(0.3s)$ (b) $\sigma_{P_{F_2}} N = 30(0.3s)$



(c) $\sigma_{f_{F_2}} N = 70(0.7s)$ (d) $\sigma_{P_{F_2}} N = 70(0.7s)$

Figure 6: ROC curves based on frequency and power for the detection of shout vs speech in noised mode