Analysis and Identification of Human Scream: Implications for Speaker Recognition

Mahesh Kumar Nandwana, John H. L. Hansen

Center of Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{mahesh.nandwana,john.hansen}@utdallas.edu

Abstract

In this paper, we present analysis of the characteristics of scream to identify its discriminating features from neutral speech. The impact of screaming on the performance of text independent speaker recognition systems has also been reported. We have observed that speaker recognition systems are not reliable when tested with scream. Also perceptual listeners test reveal that the speaker content in scream is very less for human to distinguish and classify it. This analysis will be useful for development of robust speaker recognition systems and their implementation in real-time situations.

Index Terms: speaker recognition, screaming, PMVDR, kl divergence.

10.21437/Interspeech.2014-249

1. Introduction

Mismatch between training and testing conditions is one of the primary reasons for performance degradation in speaker recognition systems. Most recent work in this area has addressed the problems of microphone, room acoustics and channel mismatch. The problem of speaker identification with non-speech sounds has not yet been addressed. Recently, several examples of non-speech based audio forensics have surfaced in USA courtroom proceedings which suggest research is needed on this topic [1, 2].

Human sounds produced via oral cavity can be classified into two broad categories: i) speech and ii) non-speech. Non-speech sounds include such vocalizations as: scream, whistle, cough, laugh, snore, sneeze, hiccups etc. This paper focuses on analysis and identification of screams which reflect a portion of the class of non-speech sounds and its implications for automatic speaker recognition.

Scream is classified as a loud vocalization in which air is passed through the vocal folds with greater force than is used in regular or close-distance vocalization. These outbursts convey alarm, surprise, displeasure or outrage, or perhaps to gain the attention of another person (or another living creature/animal). In general, scream can be classified into the following categories:

i) Fear and surprise – when someone is watching a horror movie or riding on a roller-coaster.

ii) Happiness – when someone meets a person after a long time, or a team you support wins the match.

iii) Danger and pain – when an unexpected accident happens or someone screams because of pain.

iv) Anger – screams which are result of anger or frustration fall under this category.

Earlier studies on speaker recognition systems due to changes in speech production have included speech under stress including emotions, Lombard effect and physical task stress [3, 4, 5], analysis and identification of speech in neutral and various other modes from whispered to shouted [6]. The effects of vocal efforts and speaking style on speaker verification are analyzed in [7].

Work related to scream signal processing has been considered in [8, 9, 10]. In [8], a method for real time scream detection for home applications based on a combination of continuity of log energy detection, high pitch analysis and compact MFCCs across frames using an SVM was described.

In [9], non-speech human sounds including cough, scream, laugh, and snore were classified based on multivariate adaptive regression splines (MARS) and support vector machines (SVM).

In [10], performance of acoustic features including audio spectral flatness (ASF), MFCCs, linear-prediction cepstral coefficient (LPCC) and Mel-spectrum was compared for scream detection using low power SVM classifiers.

All of these studies related to scream analysis have explored the domain of auditory analysis for audio surveillance. However, screams have not yet been studied in terms of their effect on speaker recognition systems. This is the first detailed study of analysis of non-speech scream sound along with its impact on speaker identification.

This paper is organized as follows: Sec. 2 describes the corpus collection used for the study. Next a detailed analysis of scream signals with respect to neutral speech is considered (sec. 3). In sec. 4, speaker recognition system is considered for this study, followed by a comparison of models of scream and speech in the acoustic space. Sec. 6 presents some concluding remarks and directions for future research.

2. Corpus

The corpus for this study was collected at the University of Texas at Dallas. All recordings were captured in an ASHA certified single walled sound booth at 44.1 kHz sampling rate using a table top Shure microphone.

A total of 6 male subjects from the Center for Robust Speech Systems participated in the data collection. Data collection was combination of three parts: Part 1 consists of recordings of text dependent neutral speech, comprising of 25 TIMIT sentences to be read. Part 2 consists of 12 questions to be answered for recording of spontaneous speech. In Part 3, subjects were told to scream. During recordings, the gain of the microphone was adjusted to ensure that signal strength was effective for analysis as well as to avoid clipping. The details of screaming events and speech for each subject are summarized in Table 1. Also sample clips of screaming events are available at http://crss.utdallas.edu/Projects/SID_Scream/

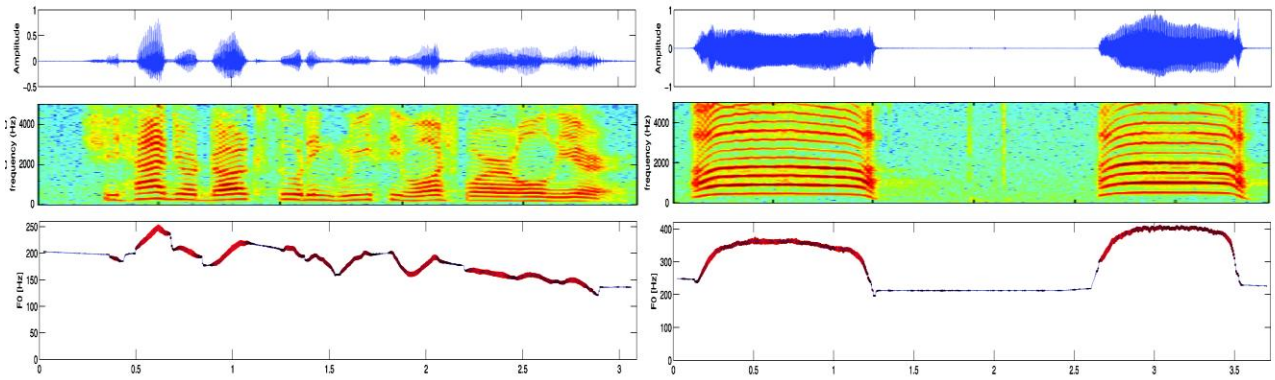

Figure 1: *Comparison of (a) time domain signal (top) (b) spectrogram (middle) (c) F0 contour (bottom) for speech (left) and scream (right) vocalization.*

It should be noted here that recordings consisted of pure scream, and should not to be confused with loud or shouted speech, etc. As mentioned in the previous section, there is no unique way to categorize a scream. Even though subjects were told in detail to imagine themselves in a particular situation, it was observed that it was hard to produce a particular type of scream naturally. So for our analysis, we will consider all types of screams under the category of scream.

Table 1. *Corpus details.*

Speaker	speech (min)	Scream events
1	26	14
2	10	14
3	19	12
4	17	13
5	14	13
6	25	13
Total	**111**	**79**

3. Analysis of Scream

This section provides a detailed analysis of changes in structure and properties of scream vocalization relative to neutral speech. For all analysis equal amounts of speech and scream data were used at 8 kHz sampling frequency. Analysis was performed in terms of: i) fundamental frequency, ii) frame energy distribution, iii) formant shift analysis, and iv) overall spectral slope. Time waveform and spectral harmonic structure changes significantly between neutral speech and scream (Fig. 1).

3.1. Fundamental frequency analysis

Fundamental frequency is one of the primary parameters found to vary under different speaking conditions. Nearly perfect F0 contours for neutral speech and scream signal were computed using [11] with an analysis window of 0.05s and a shift of 0.005s. The range of F0 was set to 50-600 Hz. F0 contours for speech and scream vocalization are shown in Fig. 1.

A drastic increase in F0 was observed, with mean and standard deviations computed across all speakers (Table 2). Scream generally results in a doubling of the mean F0 compared to neutral speech. Also, the standard deviation of F0 for scream increased significantly (2-3x increase) compared to neutral speech.

Table 2. *F0 comparison of speech and scream vocalizations across speakers.*

Speaker	Neutral Speech (Hz)		Scream (Hz)	
	Mean F0	Standard deviation	Mean F0	Standard deviation
Speaker 1	155.17	18.62	310.92	69.94
Speaker 2	157.66	28.46	262.66	80.87
Speaker 3	125.59	25.56	250.82	54.08
Speaker 4	134.88	27.11	269.45	83.80
Speaker 5	118.47	23.04	228.64	57.78
Speaker 6	162.00	33.11	242.35	91.28

3.2. Frame energy distribution analysis

Next, we consider frame energy distribution of neutral speech and scream. A K-means clustering threshold was used to remove silence frames for this particular analysis. Fig. 2 shows histograms of frame energy distributions for speech and scream vocalizations. Here, we clearly see the change in energy distribution of the signal as we move from speech to scream. The energy of frames in speech is mainly concentrated between -5 to 0 dB, whereas in scream it is in between 1 to 3 dB. In the scream vocalizations, the number of high energy frames are much greater compared to speech, and the overall histogram shifts from left to right. Thus, when vocalization moves from neutral speech to scream, frame energy results in a significant increase in voiced frames. Mean and variance of frame energy across speakers is summarized in Table 3.

Table 3. *Frame energy comparison across speakers.*

Speaker	Neutral speech(dB/frame)		Scream (dB/frame)	
	mean	Variance	mean	variance
Speaker 1	-1.7387	2.1205	1.9163	1.3018
Speaker 2	-1.0154	2.5204	2.0243	1.0545
Speaker 3	-1.0870	2.6508	2.2208	0.6152
Speaker 4	-1.9519	1.7176	2.5827	1.0583
Speaker 5	-1.5960	2.1786	1.9128	0.8928
Speaker 6	-2.3169	2.7246	1.0665	2.4488

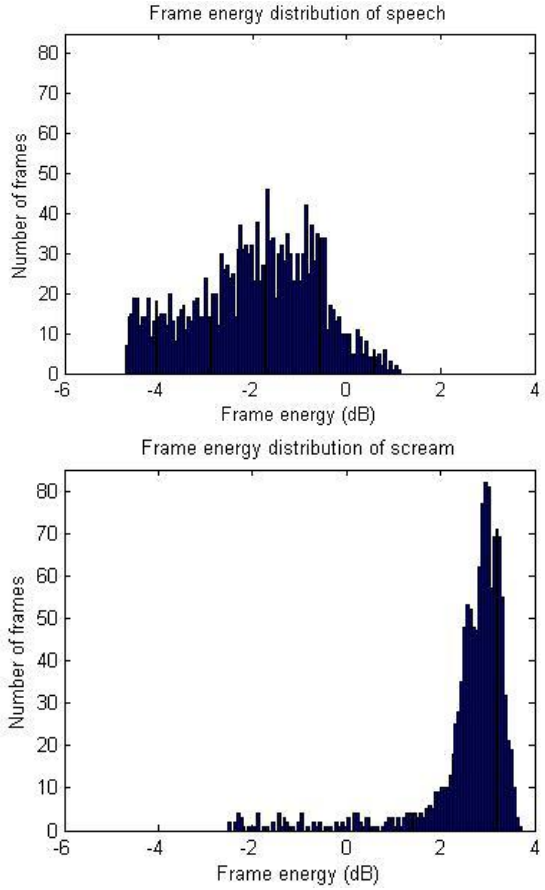

Figure 2: *Frame energy distribution of speech (top) and scream (bottom).*

3.3. Formant shift analysis

In this section, we compare the location of the first four formants for neutral speech and scream. Pure scream is analogous to the vowel /aa/. For this analysis we have use a frame of vowel /aa/ for comparison with scream for the same speaker. Formants were extracted using a 10^{th} order liner prediction analysis.

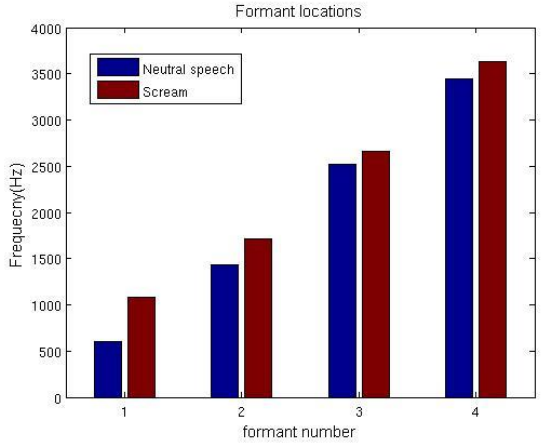

Figure 3: *Comparison of formant locations of neutral speech and scream.*

A plot of formant versus frequency is shown in Fig. 3 for speech and scream. It can be seen from the graph that there is a shift in location for every formant. However, as we increase in frequency, the higher frequency formants shift less.

3.4. Spectral slope analysis

To analyze the characteristic difference between speech and scream vocalization, we also consider spectral slope [3, 6] which generally reflects characteristics of the shape of the glottal excitation sequence, for each speaker. An energy based threshold was used to extract the potential voiced frames, with low energy and silence frames removed. A Periodogram slope was computed for the extracted voiced frames using a 256 point FFT. The resulting periodogram was averaged and a liner regression was deployed to compute the slope of the spectrum for each speaker.

Table 4. *Mean spectral slope (dB/Octave).*

Speaker	1	2	3	4	5	6
Speech	-3.44	-4.58	-6.28	-7.16	-2.90	-4.10
Scream	2.73	-0.48	-1.95	0.40	1.83	0.08

From Table 4 we clearly observe that spectral slope is steeper for neutral speech compared to scream. The change in slope for scream suggests that there are more regular shape glottal pulses in scream vocalization compared to speech, and that there is more balance between low and high frequency energy. From Fig. 1 also we also see that the spectrogram of scream is regular and sustained compared to neutral speech.

4. Performance of Speaker ID system

To analyze the effect of train/test mismatch, neutral speech and pure scream were employed in a speaker identification system. We employ a GMM-UBM based system for speaker identification. A universal background model (UBM) was constructed with all 438 male speakers of the TIMIT corpus which includes a total of 4380 sentences. A speaker specific maximum a priori (MAP) adapted Gaussian mixture model (GMM) is obtained from the UBM for each of the trained speakers [12]. The test files in trials were scored against the adapted GMM's and the resulting scores were used to obtain overall system accuracy. Performances were evaluated by computing the equal error rate (EER) for the ensemble of trials.

To observe the clear impact of screaming on the degradation of speaker recognition systems, utmost care was taken to minimize the train-test mismatch because of various other factors including channel, session, microphone, etc.

4.1. Front-end processing

In front-end processing, the data was down sampled to 8 kHz. An energy threshold and zero crossing rate based voice activity detector was used to remove silence frames. Also, speech and scream data from all speakers was windowed with a Hamming window of 20ms duration with a skip rate of 10ms. We evaluated the system performance for two acoustic features.

4.1.1. Mel-frequency cepstral coefficients (MFCC)

MFCCs are the most common features used for analysis of speech. They are computed by applying a Mel-scaled filter

bank either to the short-term FFT magnitude spectrum or to the short term LPC-based spectrum to obtain a perceptually meaningful smoothed gross spectrum.

4.1.2 Perceptual minimum variance distortionless response (PMVDR)

PMVDR feature is obtained by incorporating perceptual warping of FFT power spectrum, replacing the Mel-scaled filter bank with the minimum variance distortionless response (MVDR) spectral estimator. PMVDR based spectra have better spectral modeling ability for high pitch signals [13]. In previous studies for speaker identification, PMVDR have been shown to perform better than MFCC [14].

As we have observed, energy distributions across frames in speech and scream are very different. Therefore, to compensate for this effect in MFCC computation, we have excluded the energy coefficient C_0. A total of 36 dimensions MFCCs were calculated from C_1 to C_{12}, deltas and delta-delta. We also use 36-dimesional PMVDR features, where each feature vector contains 12 static, deltas and delta-delta. Cepstral mean and variance normalization is also applied to all features.

4.2. Open set speaker ID

For open set speaker identification, a total of 5 training tokens (10-12 sec.) are used for 32 mixtures GMM training for each speaker for different training conditions. In multistyle training equal amounts of speech and scream data was used for training. After training, the models for each speaker were adapted using MAP adaptation. A total of 3732 random trials were generated to evaluate the performance of SID systems. For scream 144 trails were used in the performance evaluation of the system. The results of open set speaker identification are summarized in Table 5.

Table 5. *Speaker verification results by train/test condition.*

Features	EER (%)	Train On		
	Test On	speech	speech+ scream	scream
MFCC	speech	17.12	18.59	51.81
	speech+scream	18.66	19.50	51.71
	scream	56.66	50.00	41.66
PMVDR	speech	14.56	14.38	51.53
	speech+scream	16.03	15.37	51.54
	scream	58.33	50.00	25.00

From Table 5 it is observed that for the case of speaker recognition from speech, PMVDR performs better than MFCC. However in the case of scream trials, systems performance decreases drastically. This result clearly confirms our earlier probe study [2] that traditional SID algorithm technology is not effective for SID on scream data; and to a secondary level that speaker ID content is significantly suppressed for automatic SID on scream.

4.3. Perceptual listeners test

Apart from automatic speaker identification systems, a human listener test was also conducted. All listeners were familiar with the subjects. In this test, a sequence of random screams were played to a listener, and each listener was asked to recognize the subject. During this task caution was exercised to ensure that the listeners did not have any prior familiarity with the speakers under screaming conditions. The highest accuracy obtained by CRSS-UTDallas listeners performing the random listener evaluation was approximately 25%.

5. Speaker Model Comparison

Finally in order to explore the acoustic spread of scream versus speech we also performed an experiment. For this we used an approximate Kullback-Leibler divergence measure. For each speaker two 32 mixtures GMM's were trained, one using speech tokens and the other using scream. 36 dimension PMVDR feature vectors were used to train both models because of their ability to model high pitch speech. Using 12 models KL divergence was computed for between class and cross class comparison of models. A confusion matrix of the comparison is shown in Fig. 4.

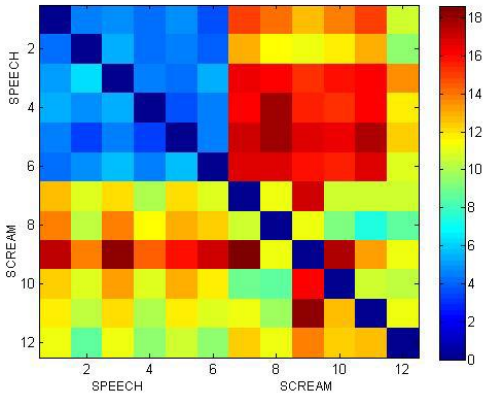

Figure 4: *Confusion matrix for speaker model comparison.*

Here, we can clearly see that the divergence between speech models is far less than the divergence between scream models and scream-speech models. This indicates that the acoustic space of speech has more phonemes and greater spread in their acoustic space.

6. Conclusions and Future Work

Unique abnormalities have been observed while analyzing human scream. It is clear that screaming results in significant changes to speech parameters, which impacts performance of speaker identification systems. Thus, this analysis suggests that existing technologies for speaker recognition are neither effective nor reliable for speaker ID using screams. Furthermore, it can also be concluded that speaker information content is suppressed in screams for human listeners as well. More effective and reliable technologies are required for speaker ID with screaming. Finally, this analysis will hopefully help in the development of robust models for speaker identification. Future research could be considered to develop compensation methods to suppress the effects of these variations in speaker recognition systems.

7. Acknowledgements

The authors would like to thank members of CRSS-UT Dallas for their help in recording and perceptual listener test.

8. References

[1] S. Skurka. “Day 6 of the Zimmerman Trial: Murder of Self-Defence? - The FBI Audio Voice Analyst”[The Huffington Post-Canada] 2 July, 2013 Available: http;//www.huffingtonpost.ca/stevenskurka/zimmerman-trial_b_3532604.html

[2] Hansen, J.H.L., and Navid Shokouhi. "Speaker Identification: Screaming, Stress and Non-Neutral Speech, is there speaker content?", IEEE SLTC Newsletter, November 2013

[3] Hansen, J.H.L., V.S.Varadarajan, "Analysis and Normalization of Lombard Speech under different types and levels of noise with application to In-Set/Out-of-Set Speaker Recognition, IEEE Trans. Audio, Speech & Language Processing, vol. 17, no. 2, pp. 366-378, Feb. 2009

[4] Hansen, J.H.L., C. Swail, A.J. South, R.K. Moore, H. Steeneken, E.J. Cupples, T. Anderson, C.R.A. Vloeberghs, I. Trancoso, P. Verlinde, "The Impact of Speech Under `Stress' on Military Speech Technology," published by NATO Research & Technology Organization RTO-TR-10, AC/323(IST)TP/5 IST/TG-01, March 2000.

[5] Godin, Keith W., and John HL Hansen. "Analysis and perception of speech under physical task stress." In INTERSPEECH, pp. 1674-1677. 2008.

[6] Zhang, Chi, and John HL Hansen. "Analysis and classification of speech mode: whispered through shouted." In INTERSPEECH, pp. 2289-2292. 2007.

[7] Shriberg, Elizabeth, Martin Graciarena, Harry Bratt, Andreas Kathol, Sachin S. Kajarekar, Huda Jameel, Colleen Richey, and Fred Goodman. "Effects of vocal effort and speaking style on text-independent speaker verification." In INTERSPEECH, pp. 609-612. 2008.

[8] Weimin Huang; Tuan-Kiang Chiew; Haizhou Li; Tian Shiang Kok; Biswas, J., "Scream detection for home applications," Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on , pp.2115-2120, 15-17 June 2010.

[9] Wen-Hung Liao; Yu-Kai Lin, "Classification of non-speech human sounds: Feature selection and snoring sound analysis," Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, pp.2695-2700, 11-14 Oct. 2009.

[10] Mak, Man-Wai; Sun-Yuan Kung, "Low-power SVM classifiers for sound event classification on mobile devices," Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp.1985-1988, 25-30 March 2012.

[11] Ewender, Thomas, Sarah Hoffmann, and Beat Pfister. "Nearly perfect detection of continuous f0 contour and frame classification for TTS synthesis." In INTERSPEECH, pp. 100-103. 2009.

[12] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." Digital signal processing 10, no. 1 (2000): 19-41.

[13] Yapanel U., Hansen, J.H.L., "A New Perceptually Motivated MVDR-Based Acoustic Front-End (PMVDR) for Robust Automatic Speech Recognition, Speech Communication, vol. 50, pp. 142-152, Jan. 2008 .

[14] Gang Liu; Yun Lei; Hansen, J.H.L., "Robust feature front-end for speaker identification," Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on , vol., no., pp.4233,4236, 25-30 March 2012

[15] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck, "MSR Identity Toolbox - A MATLAB Toolbox for Speaker Recognition Research," Microsoft Research, Conversational Systems Research Center (CSRC), October 2013.