Analysis of human scream and its impact on text-independent speaker verification

John H. L. Hansen, Mahesh Kumar Nandwana, and Navid Shokouhi

Citation: The Journal of the Acoustical Society of America **141**, 2957 (2017); doi: 10.1121/1.4979337

View online: https://doi.org/10.1121/1.4979337

View Table of Contents: https://asa.scitation.org/toc/jas/141/4

Published by the Acoustical Society of America

SUBMIT TODAY!
JASA
THE JOURNAL OF THE
ACOUSTICAL SOCIETY OF AMERICA
Special Issue: Theory and
Applications of Acoustofluidics

Analysis of human scream and its impact on text-independent speaker verification[a]

John H. L. Hansen,[b] Mahesh Kumar Nandwana, and Navid Shokouhi
Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, Texas 75080, USA

(Received 21 September 2016; revised 14 March 2017; accepted 15 March 2017; published online 27 April 2017)

Scream is defined as sustained, high-energy vocalizations that lack phonological structure. Lack of phonological structure is how scream is identified from other forms of loud vocalization, such as "yell." This study investigates the acoustic aspects of screams and addresses those that are known to prevent standard speaker identification systems from recognizing the identity of screaming speakers. It is well established that speaker variability due to changes in vocal effort and Lombard effect contribute to degraded performance in automatic speech systems (i.e., speech recognition, speaker identification, diarization, etc.). However, previous research in the general area of speaker variability has concentrated on human speech production, whereas less is known about non-speech vocalizations. The UT-NonSpeech corpus is developed here to investigate speaker verification from scream samples. This study considers a detailed analysis in terms of fundamental frequency, spectral peak shift, frame energy distribution, and spectral tilt. It is shown that traditional speaker recognition based on the Gaussian mixture models-universal background model framework is unreliable when evaluated with screams.

[http://dx.doi.org/10.1121/1.4979337]

[JFL] Pages: 2957–2967

I. INTRODUCTION

Scream, in this study, is referred to sustained, loud vocalizations with no phonological structure. Screams are a means of expressing various states of emotion including anger, distress, fear, etc. For the purposes of this and our future studies, we differentiate scream from other loud vocalizations by adding the requirement: *lack of phonological structure*. This separates scream from what we consider "yell" or "shout." Excluding phonemic content removes the implied restrictions on vocal sounds produced by the speaker in a yell, which may include one or several words.

Human sounds produced via the oral and nasal cavity can generally be classified into two broad categories: (i) speech and (ii) non-speech. Non-speech sounds include such vocalizations as: screams, coughs, whistles, laugh, snore, sneeze, hiccups, lip smack, etc., none of which contain phonological structure.[1] In screams, air passes through the vocal folds with greater force than is used in neutral speech. As far as speaker recognition is concerned, there is no clear organization or hierarchy to classify these different flavors of screams. Therefore for this investigation, we will consider all types of scream under a single category. Our focus will be on acoustic features common across all forms of scream. Figure 1 shows a sampling of the significant differences between neutral speech and scream along both time and frequency axes. The spectrogram of neutral speech shows the presence of distinct phone classes (vowels, semivowels, plosive, stop consonants, nasals, etc.). On the other hand, the spectrogram of scream shows a sustained frequency with limited, nonphonemic structure.

Mismatch between train and test conditions is the primary source of performance degradation in automatic speech systems, such as speech and speaker recognition. Unfortunately, it is virtually impossible to account for all sources of mismatch in the training data, since there are infinite possibilities of mismatch. Unlimited variability in the mismatch between train and test conditions has been proven to be a significant shortcoming of state-of-the-art speaker recognition performance Hansen and Hasan (2015). Scream is considered an important example of such mismatched cases (Huffington Post, 2013), since a standard methodology to capture scream characteristics in the training process are non-existent. However, a wide range of studies have focused on finding robust solutions to mismatch conditions including speech under stress (Hansen, 1996; Hansen *et al.*, 2000), Lombard effect (Hansen, 1988), physical task stress (Godin and Hansen, 2008), speech vocal effort (Zhang and Hansen, 2007), speaking styles (Lippmann *et al.*, 1986; Hansen, 1988; Shriberg *et al.*, 2008), long-term aging (Kelly *et al.*, 2014), and more recently non-speech sounds (Nandwana and Hansen, 2014). Such mismatch can be introduced by either speaker- or environment-dependent factors. While these factors are normally separated, there exist scenarios in which they are intertwined. For example, when environmental noise levels increase, speakers tend to alter vocal effort.

[a] Portions of this work were presented at the ISCA INTERSPEECH 2014, Singapore.
[b] Electronic mail: john.hansen@utdallas.edu

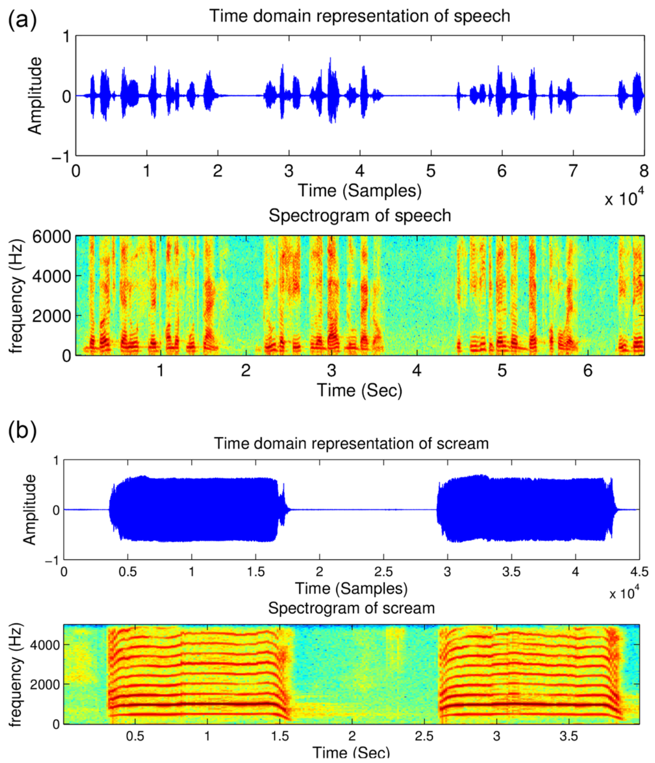

FIG. 1. (Color online) Time and F_0 domain representation of speech (top) and human scream (bottom).

Environment-dependent factors are generally extrinsic to human speech production and are due to room acoustics and recording devices. Over the past few decades, the impact of environmental factors on the performance of speaker recognition systems have been actively explored by researchers, especially in the context of the NIST SRE (Speaker Recognition Evaluation) (Multimodal Information Group, 2015). Consequently, the formulation of state-of-the-art speaker recognition systems (e.g., i-vector factor analysis, probabilistic linear discriminant analysis) has been to compensate for extrinsic sources of mismatch. However, less focus has been seen on the impact of intrinsic sources of mismatch (including scream). Recently, several examples of forensic cases have surfaced in the public eye which motivates research on this topic (Huffington Post, 2013; Nandwana and Hansen, 2014; Nandwana *et al.*, 2015). Such cases exemplify forensic scenarios in which the evidence at hand is recorded under highly intense and stressful conditions. It is reasonable to assume that a considerable amount of data in forensic trials contain stressed speech or screams, which we argue have not been significantly dealt with in the speaker recognition community.

This study addresses the following questions:

(i) What are the acoustic differences between human speech and screams? (Sec. IV).
(ii) How much does speaker verification degradation is observed in scream? (Sec. V).
(iii) What causes this performance drop? (Sec. VI).
And finally,
(iv) is there any true “speaker-dependent characteristics” in human screams? (Sec. VII).

Section II describes the background for different speaker dependent mismatches and their impact on speaker verification. In Sec. III, the UT-NonSpeech corpus developed for this study is described in detail. Section IV presents various temporal and spectral properties of acoustic signals pertaining to neutral speech and scream. In Sec. V, the impact of human scream on the performance of speaker recognition systems is illustrated. Section VI explores possible reasons for system deterioration in speaker recognition. Finally, Sec. VII presents discussion and suggests directions for future work.

II. BACKGROUND

In this study, we are interested in measuring the impact of vocalization variability within speakers on the performance of speaker recognition. Therefore, we would like to examine the signal processing aspects of various vocalization styles (which includes scream). Human screams can be considered at the intersection of speech-like sounds and non-speech vocalizations.

The forensic implications of speaker identification under extreme emotional conditions have long been a motivation for research (Saslove and Yarmey, 1980). A significant number of these studies date before many of the recent advancements in speaker recognition. Fortunately, the speaker recognition community has expressed an increased interest in exploring different speaker-dependent variabilities (e.g., Lombard, speech under stress, etc.). These are related to the levels of vocal effort used by speakers, less common in daily conversations (which we refer to as neutral speech). A number of studies have highlighted the acoustic and phonetic differences between neutral speech and speech produced with alternative vocal effort. Zhang and Hansen present a study in which five different speech modes are investigated: whisper, soft, neutral, loud, and shout (Zhang and Hansen, 2007). They analyze the differences across these modes and demonstrated the amount to which vocal effort mismatch negatively impacts speaker identification performance. In addition, the study provides a categorization of vocal effort in which high vocal effort corresponds to shouted speech and low vocal effort is associated with whisper (Zhang and Hansen, 2007). Earlier studies on speech under stress, addresses two types of vocal efforts including soft and loud (Hansen, 1988; Hansen and Clements, 1989). Those studies showed the negative impact of soft and loud spoken speech on automatic speech recognition (ASR) systems. A number of stress compensation strategies were also formulated, which resulted in significant improvement in ASR (Hansen and Clements, 1989; Hansen, 1996). Later studies considered individual speech modes such as whisper (Fan and Hansen, 2011) and shout (Hanilci *et al.*, 2013) in the context of speaker identification and proposed different compensation strategies to normalize mismatch.

Effects of various speaking styles on text-independent speaker verification have been explored in Shriberg *et al.* (2008) and Paul (1985), Chen (1987), and Hansen and Bria (1990). These studies confirm a significant challenge for systems trained on neutral speech. In Shriberg *et al.* (2008) it is noted that furtive speech introduces significant performance drop due to the mismatch with neutral speech whereas conversational and interview styles are well matched.

Among other types of vocal effort is Lombard effect; the reflexive change in speech production that occurs when speakers modify their vocal effort due to the presence of background noise in order to increase intelligibility. Over the past 20+ years, Lombard effect in speech has been analyzed in detail. Hansen and Varadarajan (2009) show that Lombard speech produced under different levels and types of noise results in different "flavors" of Lombard. Hansen and Varadarajan (2009) also illustrated the impact of Lombard effect on speaker recognition, and proposed a compensation method for train/test mismatch by adapting models trained on neutral speech with small amounts of Lombard effect speech.

Few studies have considered scream signal processing and classification (Huang *et al.*, 2010; Liao and Lin, 2009; Mak and Kung, 2012; Nandwana *et al.*, 2015). In Huang *et al.* (2010), a method for real-time scream detection using support vector machines (SVM) was prescribed for home-monitoring applications. Their detection was based on a combination of features including log-energy, high pitch analysis and compact mel-frequency cepstral coefficients (MFCCs). In Liao and Lin (2009), non-speech human sounds including cough, scream, laugh, and snore were classified based on multivariate adaptive regression splines (MARS) and SVM. Mak and Kung (2012) investigates low energy scream detection using SVM classification across several acoustic spaces: audio spectral flatness (ASF), MFCCs, linear-prediction cepstral coefficient (LPCC), and mel-spectrum.

In our earlier study (Nandwana *et al.*, 2015), we proposed an unsupervised approach for detection of human screams from continuous recordings. The proposed solution combines unsupervised vocal activity detection system and T^2-statistics. Five noise levels and five noise types were considered during the evaluation process.

These studies related to scream analysis have generally explored the domain of auditory analysis for audio surveillance. However, no study has yet considered a detailed analysis of the variations between neutral speech and human scream across different speakers, and their impact on text-independent speaker verification systems. Therefore, this represents the first detailed study to explore analysis of human scream for human vocalization, which reflects a portion of the domain of non-speech vocalizations, along with the corresponding impact on speaker recognition.

III. CORPUS DEVELOPMENT

At present, publicly, there is no corpus available for human scream research, and specifically for speaker recognition from human screams. Most prior studies on scream classification have employed sound effects for scream from movies or Internet repositories such as the *Wilhelm Scream*. While useful, such audio can and generally is corrupted by other acoustic factors such as layered background music, noise or environmental sounds, or digitizing/sample rate/audio format mismatch. In order to advance research for speaker recognition for non-speech vocalizations, two corpora were developed at the Center for Robust Speech Systems (CRSS), University of Texas at Dallas: (i) UT-NonSpeech-I and, (ii) UT-NonSpeech-II.

A. UT-NonSpeech-I

The UT-NonSpeech-I corpus contains six male speakers originally collected for scream analysis for speaker recognition. The non-speech vocalizations include scream events, and neutral speech for comparison. The experiments on UT-NonSpeech-I corpus were conducted and published in earlier studies, where details of the corpus can be found in (Nandwana and Hansen, 2014; Nandwana *et al.*, 2015).

B. UT-NonSpeech-II

The UT-NonSpeech-II corpus was developed to extend research in the area of non-speech vocalizations. Other than human screams, speaker specific cough and whistle sounds were also captured. Speaker specific cough events can be found useful for long-term health monitoring based on acoustics (Drugman *et al.*, 2013; Barry *et al.*, 2006).

A total of 56 subjects (33 males and 23 females) participated in the corpus collection. Subjects were native as well as non-native speakers of English. Each speaker's data was collected in a single session, but included a combination of three parts. Part 1 consists of neutral scripted speech, 110 phonetically balanced sentences from the IEEE recommended list of phonetically balanced sentences (Rothauser *et al.*, 1969). Part 2 contained recordings of spontaneous speech in the form of answers to six questions. For example, "summarize the story of you favorite movie," In Part 3, three non-speech vocalization events which include screams, coughs and whistles were captured for each speaker. A total of 30 min of data were recorded from each speaker. For the purposes of this study, 30 speakers (17 males and 13 females) and their entire scream and neutral speech data were used in experiments.

All recordings were captured in an ASHA certified single walled sound booth at a 44.1 kHz sampling rate using a table top Shure microphone with 16 bits per sample quantization. A multi-channel Shure pre-amp was employed with settings adjusted to avoid clipping during the A/D process for scream collection. For this corpus, a UT Dallas approved Institutional Review Board (IRB) protocol was followed during data collection. A comparison of the content of UT-NonSpeech-I and UT-NonSpeech-II corpora is shown in Table I.

TABLE I. Comparison UT-NonSpeech-I and UT-NonSpeech-II corpus across different parameters.

Corpus	UT-NonSpeech I	UT-NonSpeech II
Number of speakers	6	56
Number of males	6	33
Number of females	-	23
Read speech	TIMIT sentences (25 sentences)	IEEE recommended lists (110 sentences)
Spontaneous speech	✓	✓
Microphone	table-top	table-top
Non-speech sounds	screams	screams, coughs, whistles
English accent	non-native	native and non-native

One critical issue to overcome while recording such data was that of clipping. To address this issue during recordings, the pre-amp gain of the microphone was adjusted for each speaker to ensure that signal strength was effective for analysis as well as to avoid clipping. Gain levels were set once for each speaker to assure no level difference between scream and neutral speech. It should be noted here that recordings consisted of pure scream, and should not be confused with loud or shouted (vocal effort) speech. As mentioned in Sec. II, there is no unique way to categorize a scream. Even though subjects were told in detail to imagine themselves in a particular situation where they might scream, it is difficult to formulate a strategy to illicit screams naturally while adhering to IRB protocols. As such, the end observation during the corpus collection was that it was hard to produce a particular type of scream naturally.

IV. ACOUSTIC ANALYSIS

The effect of vocal effort on a number of acoustic, articulatory and perceptual parameters have been investigated in several studies (Pickett, 1956; Van Summers *et al.*, 1988). Most studies on vocal effort have considered shouted speech, rather than human scream. The studies on vocal effort analysis and shouted speech investigate speech under different recording circumstances, each of which tackles a particular flavor of the various forms of vocal effort. For example, Nicolaidis studies Lombard effect as a form of vocal effort by monitoring speech production generated under quiet and noisy conditions (Nicolaidis, 2012). In this case, participants would generate loud speech when speaking in noisy conditions. In other studies, deliberately loud speech has also been considered where participants were asked to produce "loud" speech (i.e., shout) during recordings (Hansen, 1988; Bond and Moore, 1990). Liénard and Benedetto use a creative method to induce vocal effort by placing microphones in three separate distances (close—0.4 m, normal—1.5 m, and far—6 m) from speakers (Liénard and Di Benedetto, 1999). Despite these elaborate analyses on shouted speech, little work has been conducted on human screams, which are extreme cases that lack the phonetic content. This section provides a detailed analysis of changes in the structure and acoustic properties of scream vocalization for humans relative to neutral speech.

In order to analyze differences in acoustic structure for these two classes, an acoustic analysis on a set of audio recordings was performed for 10 randomly chosen speakers (five males and five females) from the UT-NonSpeech-II corpus. For all analyses, equal amounts of speech and scream data were used at 8 kHz sampling frequency. Measurements were made in terms of (1) fundamental frequency; (2) frame energy distribution; (3) spectral peak shift; and (4) spectral slope. A similar probe experiment using six male speakers was previously presented by the authors Nandwana and Hansen (2014). Advancing our earlier study here, we also consider gender dependent characteristics in our analyses.

A. Fundamental frequency analysis

Fundamental frequency is one of the primary parameters found to vary under different speaking conditions. F_0 contours for neutral speech and human scream were computed from signals using the algorithm proposed in Ewender *et al.* (2009) with an analysis window of 50 and a 5 ms window shift. The F_0 range was set to 50–600 Hz. F_0 contours for speech and scream vocalizations are shown in Fig. 2. The contours for scream vary slowly across time compared to neutral speech, a variation which is expected due to the time-varying phonemic content in neutral speech.

One of the observations in increased vocal effort is an increase in the fundamental frequency (Bond and Moore, 1990). However, a main difference between other vocal efforts (e.g., Lombard) and scream lies within our definition of scream, which excludes phonemic structure. Removing this restriction from the definition of scream results in less constrained vocalizations which in turn increases the expected value of fundamental frequency over time. The average increase in fundamental frequency in loud or Lombard speech varies with the phones that are produced (Bond and Moore, 1990; Hansen, 1988). Such variation is not observed in screams, therefore the average shift in F_0 for each speaker is much more significant in screams (up to +123.5% as shown in Table II). Please refer to Hansen (1988) for a detailed analysis of F_0 variations in speech under vocal effort.

B. Frame energy distribution analysis

Here we consider frame energy distribution of neutral speech and scream. An energy threshold was used to remove silence frames for this particular analysis. Figure 3 shows histograms of frame energy distributions for speech and scream vocalizations. During recording, since the gain was adjusted for speech and scream, we have normalized the speech and scream histograms. This normalization was conducted with respect to the energy of silence frames. The energy of frames in speech is mainly concentrated between −5 and 0 dB, whereas in scream it is between 1 and 3 dB. In scream vocalizations, the number of high energy frames are much greater compared to speech, and the overall histogram shifts horizontally towards higher energies. Thus, when vocalization moves from neutral speech to scream, frame energy results in a significant increase in the number of voiced frames. Mean and variance of frame energy across speakers is summarized in Table III.

We are less interested in the increased energy levels, which state the obvious, and would like to focus more on the shape of energy distributions between speech and scream. As observed in Fig. 3, the scream energy histogram better resembles a unimodal distribution (Fig. 3, bottom), while speech energies contain multiple modes and are generated by more complex background models (Fig. 3, top). Generative modeling in speaker recognition, which includes Gaussian Mixture modeling and its various derivatives, relies on the consistency of the form of underlying models (e.g., number of mixtures, etc.) between train and test conditions. In the case of scream, the underlying model dramatically changes, which disqualifies the underlying assumptions of generative modeling used for speaker recognition.

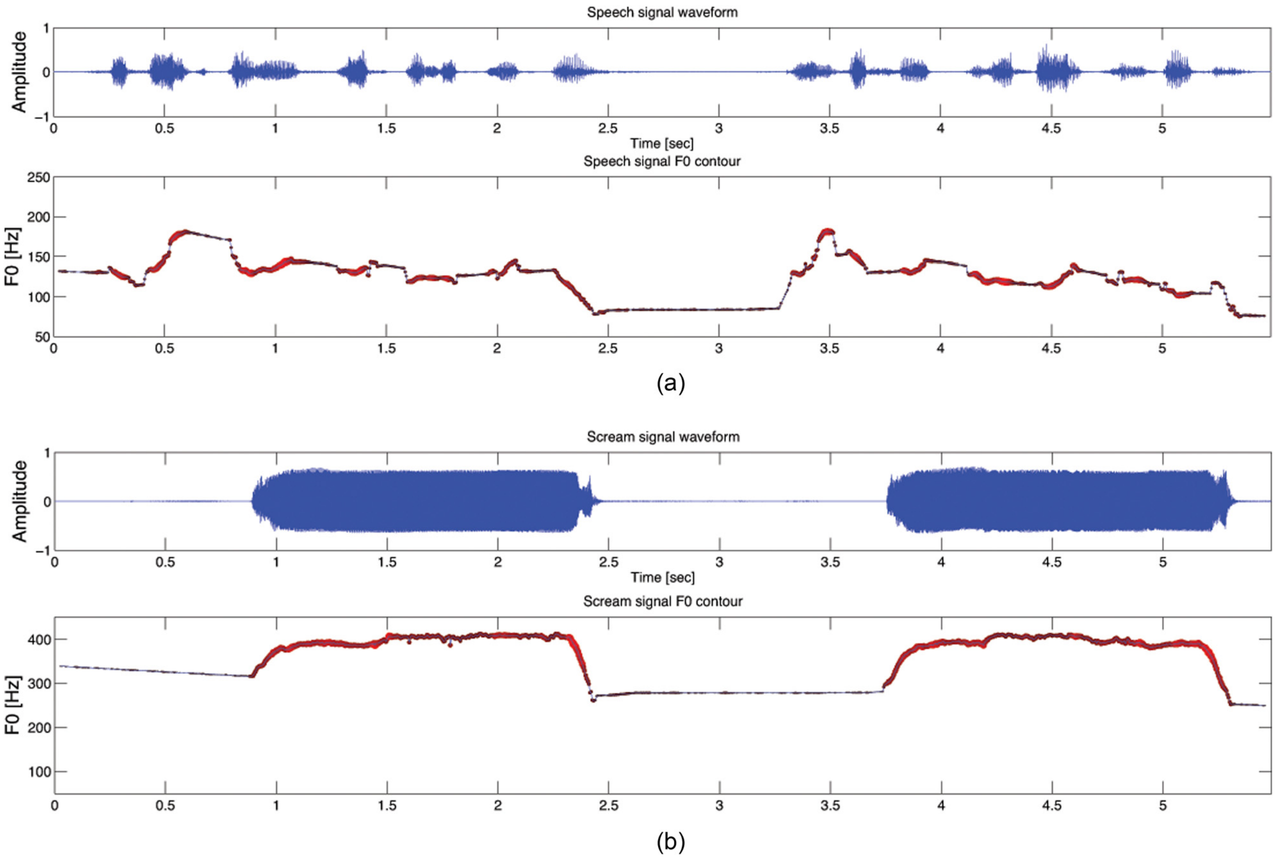

FIG. 2. (Color online) Time and frequency domain representation of speech (top) and human scream (bottom).

C. Spectral peak/formant shift analysis

In this section, we compare locations of the first three formants for neutral speech and what is believed to be vocal-tract speaker peaks for scream. In this analysis we have used PRAAT (Boersma, 2001). Formants were extracted using a 10th order linear prediction analysis for voiced frames. Plots for different formants and genders are shown in Fig. 4.

Scream is produced by pushing the tongue into the pharyngeal space (causing restriction) and therefore is expected to have a shift towards higher frequencies, particularly in F_1. Additionally, lower jaw positioning during scream results also contributes to an increase in F_1 values. The vowel space defined by F_2 and F_3 also increases in frequency. The difference between scream and speech in terms of formant frequencies is more prominent at lower frequencies (e.g., F_1).

TABLE II. F0 comparison of speech and scream vocalizations across speakers. Mean values and relative standard deviation (ratio of standard deviation to mean) are presented here. Unvoiced frames were not used in calculating these statistics.

		Male				
Speaker ID		M1	M2	M3	M4	M5
Speech	Mean F0	126.72	146.12	147.06	140.35	120.93
	RSD	5.82	5.64	5.38	6.84	6.35
Scream	Mean F0	239.25	259.93	234.75	251.29	182.11
	RSD	1.97	2.23	6.25	4.82	2.50
		Female				
Speaker ID		F1	F2	F3	F4	F5
Speech	Mean F0	167.98	157.93	201.35	158.98	223.94
	RSD	3.03	3.27	3.92	4.04	4.56
Scream	Mean F0	202.14	353.01	220.91	237.92	393.72
	RSD	1.36	1.63	1.22	1.22	8.17

D. Spectral slope analysis

To analyze the characteristic differences between speech and scream vocalization, we also consider spectral slope (Hansen, 1988; Hansen and Varadarajan, 2009; Van Summers *et al.*, 1988) which generally reflects the shape of glottal excitation sequences for each speaker. An energy based threshold was used to extract potential voiced frames. A Periodogram spectral response is produced, followed by a linear slope estimate computed for the extracted voiced frames using a 256 point fast Fourier transform (FFT). The resulting Periodogram is averaged and a linear regression between 300 Hz and 4 kHz is deployed to compute the slope of spectra for each speaker in neutral speech and scream.

From Table IV, we clearly observe that the spectral slope is steeper for all speakers in neutral speech compared to scream. The change in slope for scream suggests that there are more smooth shaped glottal pulses in scream vocalization compared to speech, and that there is more balance between low and high frequency energy. Informal visual inspections of numerous spectrograms from speech and scream suggests the same observations; the spectral slope of scream is smooth and sustained, compared to neutral speech.

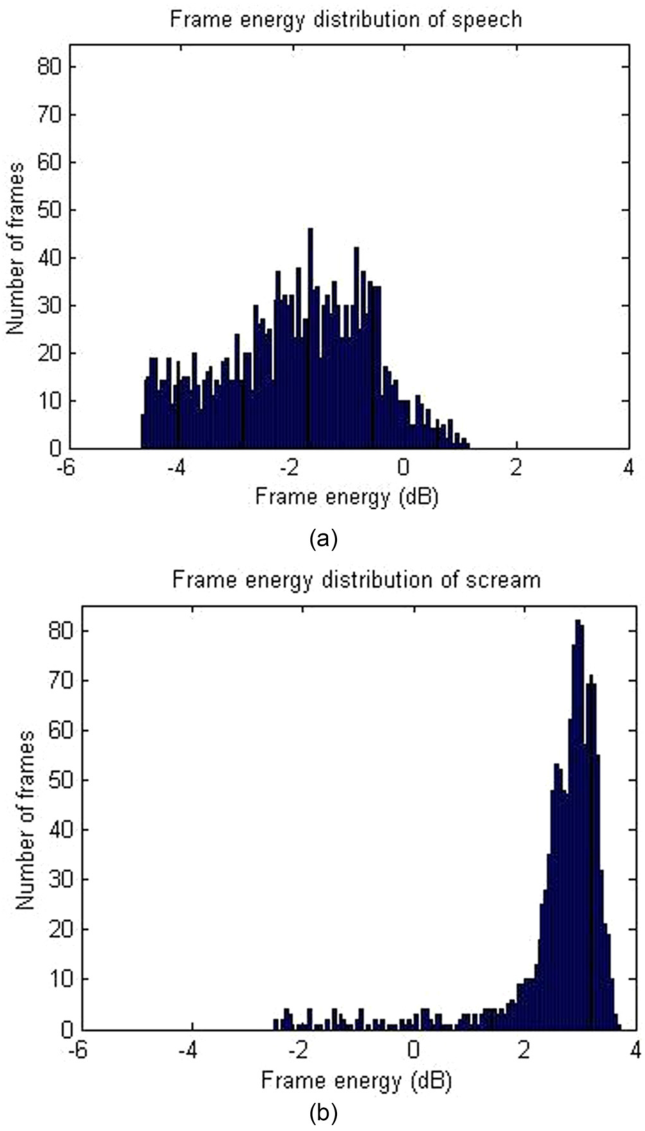

FIG. 3. (Color online) Normalized frame energy distribution of speech (top) and scream (bottom).

The acoustic analyses presented in this section depict a clear picture of the differences between scream and neutral speech. It was shown that signal energy levels are on average significantly higher in scream. In terms of spectral content, scream exhibits a shift towards higher formant frequencies. It was observed that this shift is more noticeable at lower formants (F_1) compared to higher formants (F_3 and higher). Finally, measurements show that spectral slopes are significantly more flat for screams. This flatness shows a more uniform distribution of energy across frequencies in the scream spectrum. All such changes impact the distribution of acoustic features. Section V shows how these acoustic features are vital to a speaker verification system. The dramatic change of the acoustic space from neutral speech to scream significantly affects the ability of a speaker verification system to identify the similarities between scream recordings and models trained on neutral speech (and vice versa).

TABLE III. Normalized frame energy distribution (dB/frame) across speakers.

		Male				
Speaker ID		M1	M2	M3	M4	M5
Speech	Mean	0.49	0.86	0.13	0.21	0.60
	Variance	0.76	1.20	0.69	0.86	1.00
Scream	Mean	3.13	2.72	1.20	2.33	1.65
	Variance	0.48	0.87	0.30	0.47	1.16
		Female				
Speaker ID		F1	F2	F3	F4	F5
Speech	Mean	0.04	−0.08	0.60	0.13	0.42
	Variance	0.49	0.45	1.07	0.53	0.97
Scream	Mean	2.05	2.89	2.87	2.58	3.12
	Variance	0.47	0.95	0.27	1.20	0.59

V. IMPACT ON SPEAKER VERIFICATION SYSTEM

Having pointed out the difference in speech production between neutral speech and human scream, we now shift the focus to exploring speaker recognition algorithms and their performance issues. Here, we describe our speaker verification framework used to evaluate performance on the UT-NonSpeech-II corpus.

A. Front-end processing

In front-end processing, the acoustic audio is first downsampled to 8 kHz. An energy threshold and zero-crossing-rate based voice activity detector (VAD) was used to remove silence frames. The audio signal is then pre-emphasized using an emphasizing coefficient of 0.97. Speech and scream data from all speakers was windowed with a Hamming window of 25 ms duration and a skip rate of 10 ms. We evaluated the speaker verification performance for two different types of acoustic features: (i) MFCC, and (ii) perceptual minimum variance distortionless response (PMVDR).

MFCCs are the most common features used for analysis of speech (Davis and Mermelstein, 1980). 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. Log energies are excluded, followed by a discrete cosine transform to produce a vector of coefficients for speaker verification.

PMVDR features were first proposed by Yapanel and Hansen (Yapanel and Hansen, 2008). PMVDR computes cepstral coefficients by incorporating a perceptual warping of the FFT power spectrum, replacing the mel-scaled filter bank with the minimum variance distortionless response (MVDR) spectral estimator. PMVDR does not utilize a filter-bank and instead directly warps the FFT power spectrum. These features have been shown to possess better spectral modeling of speech signals compared to more traditional feature extraction methods, particularly for mismatches caused by large variations of F_0. PMVDR has been shown to be more robust than MFCC for speaker verification tasks (Liu *et al.*, 2012). A schematic diagram of the PMVDR front-end is shown in Fig. 5.

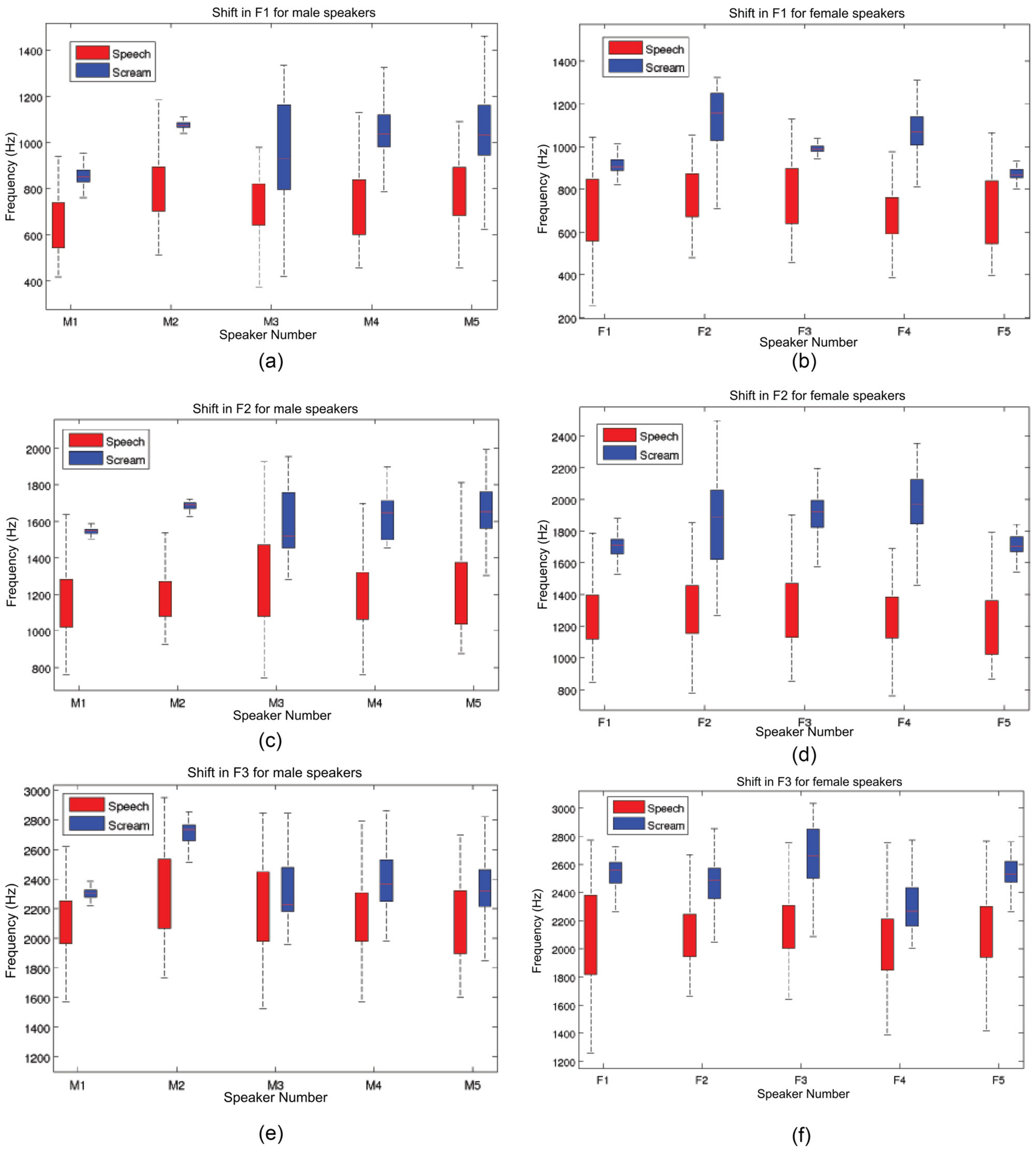

FIG. 4. (Color online) Formant shift analysis for speech and scream across speakers.

TABLE IV. Mean spectral slope(dB/octave).

	Male				
Speaker ID	M1	M2	M3	M4	M5
Speech	−4.67	−6.21	−6.42	−7.27	−5.72
Scream	−4.51	−0.66	−0.65	−4.00	−0.63
	Female				
Speaker ID	F1	F2	F3	F4	F5
Speech	−6.54	−6.04	−6.69	−5.80	−7.40
Scream	−2.55	−1.02	4.29	2.12	1.15

A total of 36 dimensional MFCCs were calculated comprised of 12 static coefficients excluding C_0, and their corresponding first- and second-order differences (Δ, $\Delta\Delta$). We also used 36-dimensional PMVDR features, where each feature vector contains 12 static, Δ and $\Delta\Delta$. cepstral mean and variance normalization was applied to all features (across scream and speech).

B. GMM-UBM framework

Our goal is to analyze the effects of train/test mismatch between neutral speech and pure scream in speaker verification.

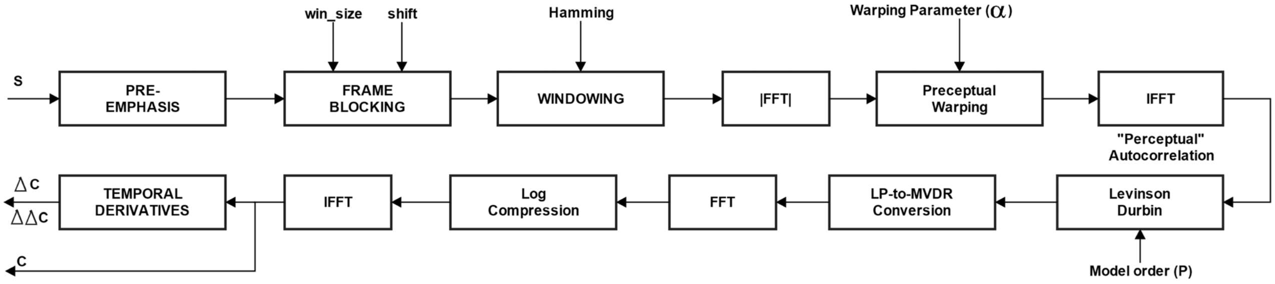

FIG. 5. A schematic diagram of PMVDR front-end.

Speakers are represented by Gaussian mixture models (GMM) derived from a Universal Background Model (UBM). The choice of a GMM-UBM framework over the state-of-the-art i-vector/PLDA setup (Garcia-Romero and Espy-Wilson, 2011) is due to the short-time nature of screams. As verified by numerous studies (Hasan *et al.*, 2013), i-vectors generated on samples with shorter duration are less reliable and result in greater performance degradation compared to a GMM-UBM system. Screams, by definition, only last up to a few seconds; hence the choice of a GMM-UBM speaker identification system.

A schematic diagram of the training and testing phase of a GMM-UBM system is shown in Fig. 6. First, A 512-mixture UBM is constructed using the expectation-maximization (EM) algorithm with features extracted from 1801 utterances from 225 native and non-native English speakers (selected from NIST SRE 04,05). Next, a speaker specific maximum *a posteriori* (MAP)-adapted GMM is obtained from the UBM for each of the trained speakers (Reynolds *et al.*, 2000) using data from UT-NonSpeech-I and II. Only GMM mean and variances were adapted in this setup. The model parameters for each speaker were estimated during the training process. Test files are scored against the adapted GMM and the resulting scores are used to calculate log-likelihood ratios. Equal error rates (EER)—where the miss rate and false-alarm rate are equal, is used as a standard measure to evaluate system performance over all trials.

To observe the clear impact of scream on performance of speaker verification, care was taken to minimize any other train/test mismatch because of various factors including channel, session, microphone, etc. Data for the UBM training was also recorded from the same microphone under similar conditions.

C. Speaker verification experiment

For speaker verification, a total of 6 training tokens (10–12 s each) were used to MAP-adapt each speaker's GMM (≈60 speakers). After training, the models for each speaker were adapted using MAP adaptation. In multi-style training (Lippmann *et al.*, 1986), equal amounts of speech and scream data (three training tokens for each class) were used. A total of ≈4000 k speech trials and ≈110 k scream trials were randomly generated to evaluate speaker verification performance. We maintained a target trials-to-impostor trial ratio of about 1:30 for both classes. The results of the speaker verification experiment for different train/test conditions are summarized in Table V.

From Table V, it is observed that for the case of speaker verification from speech and scream, PMVDR front-end performed better compared to the MFCC front-end. In the case of scream trials, system performance decreases drastically for both feature types. The performance of the baseline speaker verification system in matched neutral speech conditions could be improved by adding to the number of speaker and amount of data per speaker.

In general, speaker dependent scream data for adaptation of scream models is not available easily in real world scenarios, and so this poses a great challenge in maintaining

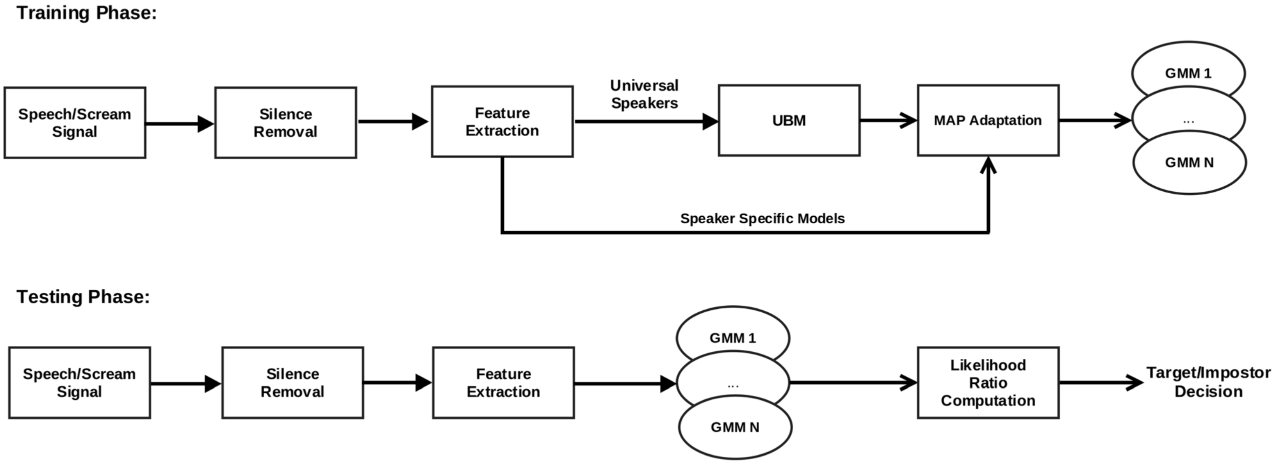

FIG. 6. Schematic diagram of training and testing phases of a GMM-UBM framework.

TABLE V. Speaker verification results for train/test conditions.

	EER(%)	Train on		
Features	Test on	Speech	Speech + Scream	Scream
MFCC	Speech	17.12	18.59	51.81
	Scream	56.66	50.00	41.66
PMVDR	Speech	14.56	14.38	51.53
	Scream	58.33	50.00	25.00

speaker identification system performance with systems trained with neutral speech. From Sec. IV and our observations of speaker verification results, traditional speaker recognition technology is not effective for speaker verification employing scream data. This stems from dramatic reductions in the feature-space span of scream data (see Secs. IV B and IV C), which reduces differences between speaker models for scream.

VI. HUMAN PERCEPTION TEST

A human listener test was conducted to investigate listeners' abilities to extract speaker identity information from scream data. Tests were designed in a speaker verification framework, where listeners were asked to determine whether two audio samples belong to the same speaker. The main test consists of two categories, (1) matching speech to scream, and (2) matching scream to scream. This categorization was made to analyze the effects of different stimuli (i.e., the first sample exposed to listeners). An additional set of speech-speech trials was also included to verify each listener's ability to perform the standard speaker verification task. A total of 10 listeners (gender balanced) participated in the test. Participants were asked to verify whether out of 20 trials presented to them (10 scream–scream and 10 scream–speech) belong to the same speaker or to two different speakers. Results show that while participants were relatively unsuccessful in both categories, matching scream to speech was more difficult and lead to 70% incorrect answers to 30% correct. On the other, detecting scream from scream samples was no different than chance (50% correct and 50% incorrect).

Listeners expressed more confidence in responding to scream–scream trials; which is in line with the results in Fig. 7. This also matches our observations from the acoustic analysis section, where we observed significant reduction in feature-space span of scream data compared to speech. When asked to describe their technique in the verification tasks, some listeners pointed out that aside from frequency characteristics they paid attention to the rise and fall and the beginning and end of each scream. Believing that most of the speaker-dependent information lies in those regions. This is interesting to us since standard speaker verification systems i-vector or GMM-UBM do not take full advantage of temporal information. They rather average out information from the entire audio to calculate likelihood ratios. This could be beneficial to future approaches for speaker verification in scream to use time trajectories as input to speaker verification systems.

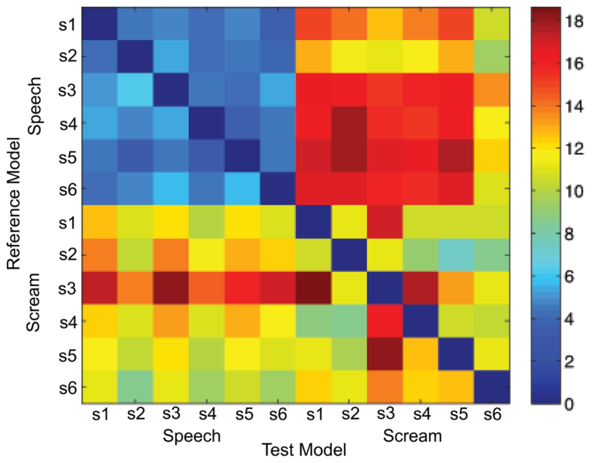

FIG. 8. (Color online) Confusion matrix for speaker model comparison. The lower the KL-divergence value (blue) the closer the two models. The KL-divergence of each model with itself is therefore equal to zero (dark blue). Warmer colors are used to represent higher KL-divergence. High KL-divergence means more separation.

VII. GMM ANALYSIS

In order to explore the differences in acoustic spaces of human scream and speech, we present an experiment to compare the similarity between their corresponding acoustic models. For this we used an approximate Kullback–Leibler (KL) divergence measure (Hershey and Olsen, 2007). Six male speakers were selected for this analysis. For each speaker, two 32 mixture GMMs were trained, one using speech training tokens and the other using scream. Thirty six dimensional PMVDR feature vectors were used to train both

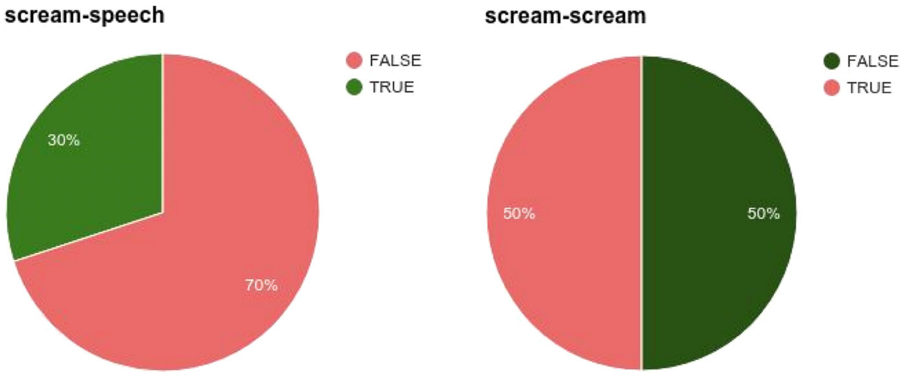

FIG. 7. (Color online) Human perception test results.

models, because of their ability to model the overall vocal tract spectral structure and less sensitivity to F_0 values. All acoustic models were compared against each other using KL divergence. A confusion matrix of the comparison is shown in Fig. 8. The diagonal represents self-comparison.

Here, we observe that between-speaker divergences are lower for speech (upper left quarter) than they are for scream (lower right quarter). Less surprisingly, the divergence between scream and speech pairs is significantly high. This indicates that the acoustic space of speech shows more consistency across speakers. From a speaker verification perspective, this is a useful feature of scream, which shows that the differences across speakers are high for scream. The downside, shown in our speaker verification experiments, is that the divergence between difference instances of scream for a single-speaker is also high, which lowers our ability to recognize speakers from previous scream samples. This is partly, due to the inefficacy of the models used to parameterize the acoustic space of scream.

VIII. CONCLUSION

In this study, we have considered a detailed analysis of human scream, a class of non-speech vocalization in order to understand both production and the presence of any potential speaker dependent traits. It was evident from the analysis that there is a significant change in many acoustic parameters which directly impacts performance of speaker recognition systems.

After establishing an analysis and understanding of the acoustic production differences, an investigation into speaker verification systems was also addressed. Due to profound differences in acoustical properties of neutral speech versus human scream, the performance of speaker verification systems trained with neutral speech degrades significantly when evaluated with non-speech vocalization such as screams.

As noted, the main purpose of this study was to understand the acoustic differences between neutral speech and human scream, and thereby incorporate this knowledge into the design of future robust speaker verification systems. The acoustic production analysis and corresponding acoustic model comparison will be helpful in developing future speech systems that are robust for speaker verification when screams are present. Future work in this domain could consider improving robustness of speaker modeling from scream. Alternative front-end features, back-end modeling strategies, multi-style acoustic modeling represent directions which are promising.

ACKNOWLEDGMENTS

This work was supported in part by the Air Force Research Laboratory under Contract No. FA8750-15-1-0205 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L.H.

[1] The focus in this study is solely on spoken English. We state this, since some non-speech sounds mentioned here, such as lip smacks, are considered valid phones in other languages.

Barry, S. J., Dane, A. D., Morice, A. H., and Walmsley, A. D. (**2006**). "The automatic recognition and counting of cough," Cough **2**, 8.

Boersma, P. (**2001**). "PRAAT, a system for doing phonetics by computer," Glot Int. **5**, 341–345.

Bond, Z. S., and Moore, T. J. (**1990**). "A note on loud and Lombard speech," in *ICSLP 1990*, pp. 969–972.

Chen, Y. (**1987**). "Cepstral domain stress compensation for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 12, pp. 717–720.

Davis, S., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech Signal Process. **28**, 357–366.

Drugman, T., Urbain, J., Bauwens, N., Chessini, R., Valderrama, C., Lebecque, P., and Dutoit, T. (**2013**). "Objective study of sensor relevance for automatic cough detection," IEEE J. Biomed. Health Inf. **17**, 699–707.

Ewender, T., Hoffmann, S., and Pfister, B. (**2009**). "Nearly perfect detection of continuous F_0 contour and frame classification for TTS synthesis," in *INTERSPEECH 2009*, pp. 100–103.

Fan, X., and Hansen, J. H. L. (**2011**). "Speaker identification within whispered speech audio streams," IEEE Trans. Audio, Speech, Lang. Process. **19**, 1408–1421.

Garcia-Romero, D., and Espy-Wilson, C. Y. (**2011**). "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, pp. 249–252.

Godin, K. W., and Hansen, J. H. L. (**2008**). "Analysis and perception of speech under physical task stress," in *INTERSPEECH 2008*, pp. 1674–1677.

Hanilci, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., and Ertas, F. (**2013**). "Speaker identification from shouted speech: Analysis and compensation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 8027–8031.

Hansen, J. H. L. (**1988**). "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, p. 47.

Hansen, J. H. L. (**1996**). "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," Speech Commun. **20**, 151–173.

Hansen, J. H. L., and Bria, O. N. (**1990**). "Lombard effect compensation for robust automatic speech recognition in noise," in *ICSLP 1990*, pp. 1125–1128.

Hansen, J. H. L., and Clements, M. A. (**1989**). "Stress compensation and noise reduction algorithms for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 266–269.

Hansen, J. H. L., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: A tutorial review," IEEE Signal Process. Mag. **32**, 74–99.

Hansen, J. H. L., Swail, C., South, A. J., Moore, R. K., Steeneken, H., Cupples, E. J., Anderson, T., Vloeberghs, C. R., Trancoso, I., and Verlinde, P. (**2000**). "The impact of speech under 'stress' on military speech technology," NATO Res. Technol. Org. RTO-TR-10, AC/323 (IST) TP/5 IST/TG-01.

Hansen, J. H. L., and Varadarajan, V. (**2009**). "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," IEEE Trans. Audio, Speech, Lang. Process. **17**, 366–378.

Hasan, T., Saeidi, R., Hansen, J. H. L., and van Leeuwen, D. A. (**2013**). "Duration mismatch compensation for i-vector based speaker recognition systems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7663–7667.

Hershey, J. R., and Olsen, P. A. (**2007**). "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4, pp. IV–317.

Huang, W., Chiew, T. K., Li, H., Kok, T. S., and Biswas, J. (**2010**). "Scream detection for home applications," in *IEEE Conf. on Industrial Electronics and Applications*, pp. 2115–2120.

Huffington Post (**2013**). "Day 6 of the Zimmerman trial: Murder or self-defence?," http://www.huffingtonpost.ca/steven-skurka/zimmerman-trial_b_3532604.html (Date last viewed 5/29/2015).

Kelly, F., Saeidi, R., Harte, N., and van Leeuwen, D. (**2014**). "Effect of long-term ageing on i-vector speaker verification," in *INTERSPEECH 2014*, pp. 86–90.

Liao, W.-H., and Lin, Y.-K. (**2009**). "Classification of non-speech human sounds: Feature selection and snoring sound analysis," in *IEEE Int. Conf. on Systems, Man and Cybernetics*, pp. 2695–2700.

Liénard, J.-S., and Di Benedetto, M.-G. (**1999**). "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411–422.

Lippmann, R. P., Mack, M., and Paul, D. (**1986**). "Multi-style training for robust speech recognition under stress," J. Acoust. Soc. Am. **79**, S95.

Liu, G., Lei, Y., and Hansen, J. H. L. (**2012**). "Robust feature front-end for speaker identification," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 4233–4236.

Mak, M.-W., and Kung, S.-Y. (**2012**). "Low-power SVM classifiers for sound event classification on mobile devices," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1985–1988.

Multimodal Information Group (**2015**). "NIST Speaker Recognition Evaluation (SRE)," http://www.nist.gov/itl/iad/mig/sre.cfm (Date last viewed 5/29/2015).

Nandwana, M. K., and Hansen, J. H. L. (**2014**). "Analysis and identification of human scream: Implications for speaker recognition," in *INTERSPEECH 2014*, pp. 2253–2257.

Nandwana, M. K., Ziaei, A., and Hansen, J. H. L. (**2015**). "Robust unsupervised detection of human screams in noisy acoustic environments," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 161–165.

Nicolaidis, K. (**2012**). "Consonant production in Greek Lombard speech: An electropalatographic study," Riv. Linguist. **24.1**, 65–101.

Paul, D. B. (**1985**). "Training of HMM recognizers by simulated annealing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 10, pp. 13–16.

Pickett, J. M. (**1956**). "Effects of vocal force on the intelligibility of speech sounds," J. Acoust. Soc. Am. **28**, 902–905.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (**2000**). "Speaker verification using adapted Gaussian mixture models," Digital Signal Process. **10**, 19–41.

Rothauser, E., Chapman, W., Guttman, N., Nordby, K., Silbiger, H., Urbanek, G., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Saslove, H., and Yarmey, A. D. (**1980**). "Long-term auditory memory: Speaker identification," J. Appl. Psychol. **65**, 111–116.

Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S. S., Jameel, H., Richey, C., and Goodman, F. (**2008**). "Effects of vocal effort and speaking style on text-independent speaker verification," in *INTERSPEECH 2008*, pp. 609–612.

Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**, 917–928.

Yapanel, U. H., and Hansen, J. H. L. (**2008**). "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," Speech Commun. **50**, 142–152.

Zhang, C., and Hansen, J. H. L. (**2007**). "Analysis and classification of speech mode: Whispered through shouted," in *INTERSPEECH 2007*, pp. 2289–2292.