Profiling Hoax Callers

Rita Singh [†], Joseph Keshet [‡], Eduard Hovy [†‡]
[†]Computer Science Department, Carnegie Mellon University, Qatar
[‡]Computer Science Department, Bar Ilan University, Israel
[†‡]Language Technologies Institute, Carnegie Mellon University, USA

***Abstract*—Hoax calls annually cost law enforcement and security agencies over a billion dollars, and sometimes lives. Bogus bomb threats, "swatting" calls to the police, hoax calls to the coast guard etc. cause these agencies to respond, deploying personnel and resources needlessly. The response itself could cause direct danger to innocent citizens, while also drawing resources away from genuine emergencies that could otherwise have been expeditiously attended to. Law enforcement agencies would hence benefit greatly from technologies that could assist them in identifying the circumstances of a hoax call, or identifying the hoax callers themselves. These could lead to more informed responses to hoax calls, or to the arrest of the perpetrators. In this paper we describe technologies to *profile* hoax callers. Profiling in this context refers to the estimation of the speaker's personal traits, and of their physical surroundings, from their voice. This is a difficult task, particularly because the hoax calls are often very short, degraded in audio quality, and highly dramatized, with the callers attempting to disguise their voices to prevent identification. We present aspects of current technology in various fields that we have applied to this problem, and the challenges that remain in formulating reliable solutions.**

I. Introduction

Hoax calls constitute a special category of voice-based crimes that are committed to elicit serious, costly and sometimes deadly responses from security, emergency and law-enforcement agencies. The people who make these calls may have different motivations, ranging from mental issues such as malicious pleasure and societal hatred to carefully planned decoy activities, but in most cases they are fully aware of the consequences of their actions, and the dangers posed to the community. Hoax calls are largely anonymous calls, and the perpetrators take care to not reveal any clue of their identity. In these voice-based crimes, the recorded voice of the speaker is often the only available evidence.

This paper addresses the problem of using such voice evidence to generate a detailed description, or a *profile*, of the speaker through analysis of the speaker's voice. The information derived falls under two broad categories: a) Bio-relevant parameters, or *bio-parameters*, of the speaker and b) Environmental parameters. The set of **bio-relevant paramters** includes, but is not limited to *Physical parameters*: height, weight, body-shape, facial structure; *Physiological parameters*: age, presence or absence of medications in the body; *Behavioral parameters*: dominance, leadership; *Medical parameters*: state of the speaker's physcial and mental health, presence of particular disease(s) and disabilities; *Demographic parameters*: race, geographical origins, level of education; *Sociological parameters*: social status, income etc. The set of **environmental parameters** includes details of the location of the speaker, objects surrounding the person, and the devices used by the person at the time of speaking.

Prior studies in multiple fields have shown that human voice carries traces of the speaker's bio-parameters. The general, as yet unproven, hypothesis is that voice carries the biomarkers of *every* parameter that influences the speaker's biological system and psychological state. In addition, once the voice signal is captured, traces and influences of the speaker's environment are also captured alongwith it. This paper is intended to be a brief exposition of the mechanisms for profiling, i.e. for the identification, extraction and interpretation of these traces, that we have devised lately, and of the scientific challenges that remain to be addressed.

II. The complexity of hoax calls

In contrast to clean voice signals captured in studio-quality environments, even the shortest hoax calls comprising just one word such as "Help" or "Mayday" can be extremely complex in the spectral domain. There are two sources of this complexity. One is introduced by the vocal maneuvers of hoax callers in attempts to simulate panic or hide their identity (or both), and the second relates to losses incurred in the process of recording and transmission of the acoustic signal. In this section we explain the nature of these complexities briefly.

A. Vocal maneuvers in hoax calls

It is our (interesting) observation from a study of many real cases that hoax callers often modify their natural voice, sometimes in extreme ways. Although probably not aware of the biometric potential of their voice, they instinctively attempt to hide their identity by disguising it. In some cases such as false bomb threats and public safety related calls, they may use external devices that distort their voice but as long as the voice remains intelligible to the intended victim(s), the hoax callers achieve their goals. In other cases they try to sound like a real (albeit fictitious) person *other than themselves*. This is because their goal is to induce the emergency services to respond, and this may not be achieved if the voice quality itself causes the response agencies to flag it as a hoax.

In a real emergency situation, the caller may be in physical danger, and under extreme stress. The caller may be screaming, shouting, crying or expressing extreme emotion, and these factors result in highly distorted voice signals. In a hoax call, the distress may be simulated. Nonetheless, the effect is the same on the voice signal in terms of the degree of distortion.

Fig. 1 shows an example where the same person made hoax calls on two different occasions, pretending to be a curious observer in one after having made a hoax call, and simulating panic in his voice in the other. The spectral characteristics observed are very different in each case. Profiling algorithms must be agnostic to such high degrees of variation in the human voice. The problem is that so far, there is no clear thesis on the variations of human voice and their characteristic properties. Observation of several pairs of such calls reveals that different people may simulate panic in different ways. We are in the process of building our understanding of this phenomenon.

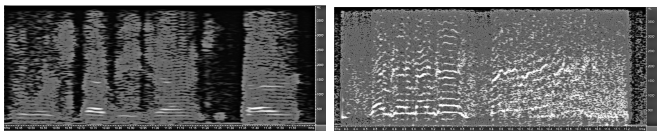

Fig. 1. **Left**: Narrowband spectrogram of a hoax caller speaking normally. The spectrogram shows the logarithm of the energy at each frequency as a function of time as a grey-scale map, with a cutoff at 100dB below peak energy in any frame for clarity. **Right**: The word "Mayday" spoken five times with increasing (fake) degree of panic by the same person in a highly noisy transmission. The grey patches show noise in the signal.

B. Device and channel induced losses

Hoax calls are almost entirely wireless communications that are transmitted over radio, telephone or internet channels. In real emergencies, distress transmissions are made from more than normally disturbed circumstances, and often contain high levels of noise. In general, the audio quality of hoax calls is also highly variable and poor. The callers are neither discerning about the location of their call (which may be noisy), nor of the devices they use to capture and transmit their voice, which may be of poor quality. The transmission channels used also employ lossy compression and coding/decoding schemes, which are designed to roughly preserve the perceptual quality of speech, but not necessarily the fidelity at all frequencies. In addition, there may be channel distortions and other attenuations introduced due to bad microphones, improper settings, bad handling, reverberation and echo in the environment etc. All of these factors cause the profiling-relevant information in the acoustic signal to become more obscure and difficult to extract reliably.

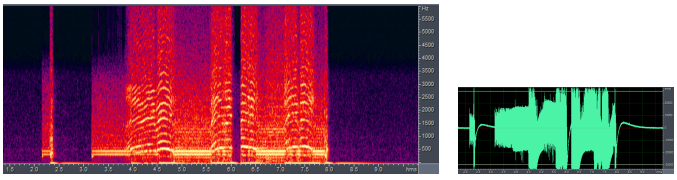

Fig. 2. **Left**: Spectral distortions introduced due to the transmission channel and devices used. Aliasing severely obscures the information in the speech regions. Data dropouts appear as vertical gaps in the spectrogram. **Right**: Coding induced uncorrected amplitude modulation in the time domain signal for the same recording.

The example in Fig. 2 is that of an unknown caller's voice transmitted over a VHF channel. It shows many different key issues that appear in real hoax calls. Firstly, information above 4kHz is lost due to the coding scheme used. Information below 240Hz is completely cut off. In the speech regions of this signal, the frequency aliasing introduced due to clipping in the time domain smears the spectral patterns into regions where there is really no information (below 240Hz and above 4kHz), and also obscures much of the spectral patterns in between. Note that the human vocal tract produces frequencies between approximately 50 and 6800 Hz during normal speech. The pitch or fundamental frequency of speech is around 120Hz for men and 210Hz for women. This is completely removed by the 240Hz cutoff during transmission. In fact, the cutoff range of 240-4000Hz also removes information in higher frequencies that is necessary for the disambiguation of many speech sounds, and diminishes their potential for use in profiling. Profiling must then be done using the remaining information, which is attenuated *differently* for male and female voices due to the inherent differences between them [1], [2].

Channel noise is visible in Fig. 2 as a combination of largely diffuse random noise and clearly visible (almost) pure-tone spectral components that are harmonics of a 123Hz tone, presumably due to the lack of suppression of CTCSS (Continuous Tone Coded Squelch System) tones used by amateur VHF and UHF equipment [3]. Note also the significant amplitude modulation introduced in the signal waveform due to coding issues. A closer study of hoax calls recordings reveals the presence of several other types of problems such as data dropouts, spectral holes, spectral smearing etc.

The challenge here is that current techniques that compensate for these effects, introduce new artifacts that degrade the speech signal itself. These techniques therefore cannot be used for profiling as-is, and must be modified significantly.

III. Deducing bio-relevant parameters

Techniques that are successfully applied to do speaker matching for speaker identification and verification are not suitable for use in profiling for many reasons. Profiling is not a problem of identification, verification or matching where prior voice templates are available for comparison. Profiling must often be done without the possibility of comparison – a hoax call is received and we must generate a description of the speaker from it by a direct analysis of the acoustic signal. This of course does not exclude the possibility of *comparative profiling*, where we use an existing database of voices from people with known biometric parameters, and assign the profile of the closest matching person to the hoax caller.

A. Micro-articulometry

Our current approach to profiling falls under an area that we designate as *micro-articulometry*. The term *articulometry* itself conventionally refers to the measurement of the movements, dimensions and positions of the articulators in the human vocal tract during the process of speech production. We use it to refer to the measurement of *micro* properties of automatically extracted articulatory-phonetic units of speech. In other words,

the term refers to the fragmenting of speech into its consistent compositional units, and measurement of their properties at extremely fine levels in time, frequency and other domains.

In micro-articulometry, we measure these features in a manner that they capture localized and consistently exhibited characteristics of phonemes. These are typically the central cores of each phoneme, since in continuous speech, the spectral patterns at the extremities of each phoneme may be modified by those of the previous and succeeding phoneme [4]. These representative sub-phonetic regions are automatically extracted using a state-of-art Hidden Markov Model based speech recognition system that is specially trained with entropic constraints to extract accurate *sub-phonetic* segmentations [5]. In a variant of this, we also extract micro-features from the transition regions between combinations of adjacent phonemes, e.g. in a single-word "Mayday" call, we may extract the rise of pitch in the transition between [M EY], and [D EY], rather than from only the central regions of the sounds represented by M, D and EY (we use capital-letter sysmbols to denote phonemes throughout this paper). In the paragraphs below, we give some examples of micro-features.

Fig. 3 shows an example of a micro-feature in the *frequency* and *time-frequency* domains. The top panel of this figure is the narrow-band spectrogram of a rendition of Popeye the Sailor Man [6] by Jack Mercer. This was recorded in 1935, at a time when there were few audio processing techniques available to artificially render the unusual spectral characteristics seen in the spectrogram. The voice actor had to actually produce the sounds as seen in this figure. The spectrogram shows *frequency modulation* on each harmonic clearly. This can be measured via frequency demodulation techniques and is an example of a micro-feature. In hoax calls that simulate panic, this kind of modulation is often present. The lower panel of Fig. 3 shows a micro-feature in the time-frequency domain. This is the *bandwidth* of each *harmonic*, and is outlined by fine black contours (clearly visible on enlargement on-screen). We see that the bandwidths are different for each harmonic. In our experience, the bandwidths are very characteristic of each speaker. We continue to investigate the full potential of this new micro-feature.

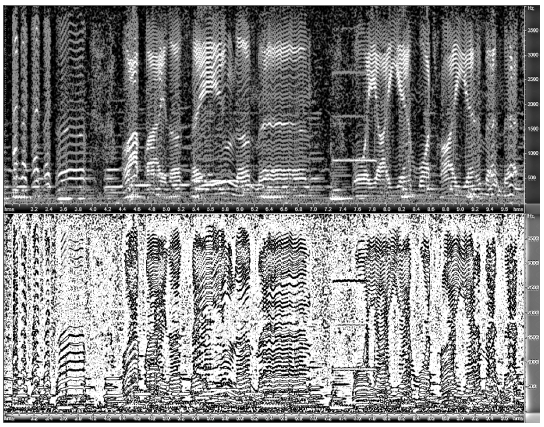

Fig. 3. **(a)** Voice signal showing several micro characteristics

Fig. 4 shows a micro-feature in the *time* domain: the Voicing Onset Time (VOT). This is relevant in phoneme combinations of a plosive sound, such as T, P, D, B, G, K followed by a voiced sound, such as a vowel. In the production of a plosive, the vocal tract is closed at some location (such as the lips for a P, and the palate for a K) and air pressure builds up behind the location. This is then suddenly released, and the articulators move to the configuration for the next sound. During the stop and release phase of an unvoiced plosive sound, the vocal folds do not vibrate. If the next sound is voiced, there is a time gap between when the vocal folds are at rest, and when they begin vibrating. This is usually in the order of milliseconds, and is

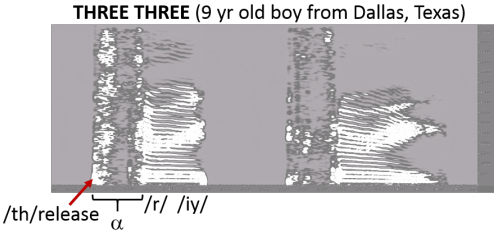

Fig. 4. Two instances of the word THREE spoken by a 9 year old boy from Texas (source of data: [7]). Although the durations of the words are different, the voicing onset time α remains the same. In practice, the VOT changes within a small range around the mean for each speaker.

highly speaker dependent since the key factor that controls this time interval is the inertia of the muscles that control the speaker's vocal folds. The speaker, even in the most extreme forms of voice disguise, does not have voluntary control over this inertia. For profiling, we derive a large number of other such micro-features, which we do not list explicitly here. The algorithms used to derive them are very specific to the feature extracted. In each case, *high-accuracy* measurements of these features is critical to the performance of prediction algorithms. As a result, in addition to identifying features, significant effort is also required to devise algorithms that can measure them accurately. We use sophisticated algorithms based on high-accuracy spectral analysis [8] and structured prediction [9] to derive such micro-features, and continue to devise newer, more accurate algorithms for their estimation.

1) Parameter prediction with microfeatures: Once the features are derived, we use them to first identify the set of articulatory-phonetic units that best predict the profile parameter to be estimated from those features, and subsequently use this set to derive the final estimate of the parameter. We explain this procedure in greater detail below.

Identifying the most predictive units: For the same parameter, the most predictive phonemes may be different for different micro-features derived from the signal. These are therefore learned separately for each micro-feature type. For example, for the estimation of height, using high-resolution robust linear cepstra as features and the publicly available TIMIT [10] database, we find that the most predictive units from a standard speech database are the vowels EY, AE and IY. This result is shown in Fig. 5, and the procedure used to obtain it, is outlined in [5].

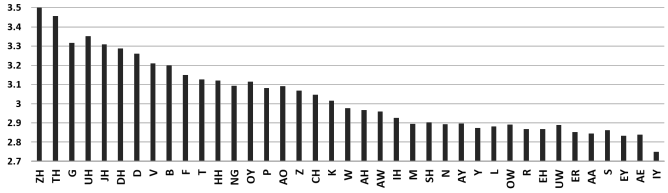

Fig. 5. RMS error in inches of the predicted values of height for each phoneme. This figure is reproduced from [5].

Our strategy takes into account the fact that the various speaker parameters we wish to derive are not independent of one another, and what we observe in the voice signal may be the *joint* effect of many different parameters. Moreover, the relationships are often not linear, and cannot be well quantified through correlation analysis (which assume an underlying linear relationship). We therefore employ an alternate strategy to characterize the potentially non-linear statistical relationships between acoustic features and body parameters [5]. We train a non-parametric *predictor* that attempts to predict the target parameter from collections of features. The predictor is optimized though appropriate cross-validation procedures to minimize over-fitting to the data. We then predict the target parameter for a held-out data set. The correlation between the predicted and true values of these parameters provides quantitative evidence of the relationship between the acoustic features and the predicted parameter. Based on these measurements, we select the most predictive phoneme for the parameter, given the feature type under consideration.

Estimating the profile parameters: To estimate a given parameter, we use the same non-parametric models as in the training stage, and the same corresponding features derived from the segmented hoax call data, to generate predictions from the set of most predictive units identified in the previous step. Following this we use effective fusion strategies [11] which we have recently shown to work well in multimedia retrieval using audio tracks. In some cases, we average the predictions obtained from all the instances to obtain a single phoneme-specific prediction for the subject. The estimates from all phonemes can also be combined using an inverse R^2 weighted interpolation to obtain a single aggregate prediction for the parameter. To comply with the Daubert criteria mentioned in Section V-A, we generate confidence values with our estimates. Currently, we use the statistical correlations observed between the parameter being estimated and the corresponding features, to generate the confidence values.

B. Alternative approaches

Neural network approaches have been shown to be able to automatically extract relevant features without resorting to preconceived ideas of what feature(s) or feature-type(s) may be important. This fact has lately been leveraged in various tasks such as speech recognition and computer vision. They have not been widely used for biometric characterizations from voice primarily because they require large amounts of data to train. Once such data are available, various neural network formalisms can be investigated to directly learn relationships between the basic spectral and other characterizations of the speech signal and the relevant body parameters.

IV. Deducing the speaker's physical environment

While each articulatory phonetic unit of speech is affected differently by different bio-relevant parameters of the speaker, the effect of environmental parameters is largely uniform on the all the units. The problem of deriving environmental factors from voice (or the acoustic signal it is embedded in) is therefore not that of micro-articulometry, but of conventional acoustic event and object detection. Accordingly, the features we select for this part of the profiling are both micro- *and* macro-level features. The techniques used largely do not require the segmentation of speech into phonemes.

The physical environment can affect voice in two ways: by affecting the *human* and causing changes in the voice production process, and by causing changes in the voice *after* it is produced. In this section we focus on the latter category of changes, wherein the influence of the environment can either be *active*, where new sounds from the environment additively superimpose on the recorded voice signal, or it may be *passive*, where the objects in the environment (or its state, such as temperature), modify the voice signal. We describe these influences briefly below.

A. Active elements profiling (AEP)

Active elements, or active environmental factors, are those aspects of the environment that actively emit signals that influence a sound recoding. Active factors are both natural and man-made sound-emitting factors, such as traffic, wind, birds, fans, air-conditioners etc., and signal modifying factors such as Electric Network frequency (ENF) variations and other electromagnetic disturbances that get recorded alongwith the speaker's voice. In most cases, these have characteristic sound patterns or *signatures* that combine additively with a speaker's voice, and also get recorded alongwith the hoax caller's voice. Active factors are deduced by extracting the signatures of these objects from the voice recording, and matching them against the signatures of known objects, e.g. for the hoax calls received from waterbound vessels, we study the signatures of boat engines, helicopters and other maritime sound-emitting objects. Our team collaborates directly with the U.S. Coast Guard Research and Development Center and Coast Guard Investigative Service to better understand typical maritime environmental challenges.

Recognizing the signatures of sound emitting objects: The best techniques for isolating the signatures of sound emitting objects for speech signals are based on NMF-based signal separation [12]–[14]. These techniques however work best when examples of the sounds expected to be separated are available. They also work best only for clean signals. Where feasible, we use these techniques to separate the speech from other sounds for identification. Where they are not feasible, and for identification itself, we use bag-of-words based audio event detection techniques [15]. These have been successful in multiple contexts in multimedia applications.

For identification, our bag-of-words based object classifiers are built using the sound examples collected under our Never Ending Learning of Sound (NELS) project at Carnegie Mellon University, which scours the web to automatedly identify and collect examples of various sound categories. In most cases, the actual detection of background sounds can be well performed by simple Support Vector Machine classifiers. Often, multiple sound emitting objects are present in the speaker's environment, and these must be disambiguated during the classification process. We continue to work on this problem.

Locating speakers through ENF analysis: In USA, the electric network frequency (ENF) is 60Hz. However, this is not constant – depending on the overall electrical power load, the actual frequency in fact varies continuously and randomly between ∼57 and ∼63 Hz. These ENF variations are the same across *the entire grid* and are constantly recorded by various agencies. The ENF variations from the power lines around get embedded in any audio signal that is generated or recorded by a device that is plugged into a wall outlet, and although very faint, can be extracted from the recorded signal. The sequence of frequency patterns over the period of the recording are often distinctive enough that a comparison with the grid ENF fluctuations on record can reveal the exact time at which this sequence occurred. ENF-based deductions have been used to solve several cases in the UK, where the entire nation is on a single grid. Even if there are multiple grids (as in the USA), ENF patterns are sufficiently unique that both the grid and the time can be determined. This can help pinpoint the location of the recording to the region of a particular grid, and time of the recording to within seconds. Also, the *absence* of ENF signatures in the recording could mean that the device that made the transmission was not plugged into a grid (as in a vessel at sea or on the river, or a battery-operated device.

Note that active factors can sometimes modify the manner in which humans produce voice [16], especially when the environmental noise level exceeds 55dB SPL. We do not address these changes in this paper, as the techniques we would use in these situations would be the same as those used for voice disguise.

B. Passive elements profiling (PEP)

Passive factors include transmission devices and channels that can modify, attenuate and distort sound in characteristic ways, but unlike ENF, without adding more information to them. Passive factors also include objects that cause the reflection, refraction, diffraction, reverberation or echo of sound.

Deducing the speaker's physical soundings through reverberation analysis Reverberation occurs when the source of an acoustic signal is in the vicinity of surfaces that reflect sound, such as walls in a room, glass panes etc. The sound gets multiply reflected from these surfaces. In recorded signals, these multiple reflections appear as delayed copies of the original signal that are added on to the original signal.

In a recorded voice signal, reverberation causes *spectral smearing*, where every frequency is extended in time beyond the end of its production by the source. This is shown in the left panel in Fig. 6, where the horizontal extensions from each harmonic of the spoken word are the spectral smear associated with it.

TABLE I
ABSORPTION COEFFICIENTS FOR SOME CONSTRUCTION MATERIALS

Materials	**Coefficients for (Hz or CPS)**					
	125	250	500	1000	2000	4000
Heavy glass	.18	.06	.04	.03	.02	.02
Ordinary glass	.35	.25	.18	.12	.07	.04
Concrete, Terrazzo, marble or glazed tile	.01	.01	.02	.02	.02	.02

The right panel in Fig. 6 shows a three-dimensional spectrographic visualization of the estimated "room impulse response" of a typical room. It shows the magnitude of the response at all frequencies between 0-8000Hz as a function of time, in response to a hypothetical impulsive sound at time $t = 0$. As we can see, the impulse response extends across all frequencies with occasional impulsive peaks. The locations of the peaks depend on the dimensions of the recording space. The shape of the response across frequencies and the *reverberation time* – the time taken for the response to fall by 60dB – depend on the dimensions of the room and the material composition of the reflecting surfaces. Analysis of the room impulse response can hence reveal information about the materials and dimensions of the recording enclosure [17].

The extraction of the actual response of the recording space from a recording, however, remains an unsolved problem. *Approximations* to it can nevertheless be obtained through techniques such as non-negative spectral matrix factorization [18]. The extracted room impulse response can be used to estimate peaks in the reverberation, the frequency response, and the reverberation time constant at different frequencies. Combined with the known reflective properties of common construction materials, e.g. as in (Table I), these enable us to guess the dimensions and material composition of the speaker's physical surroundings. Absence of reverberation indicates open space.

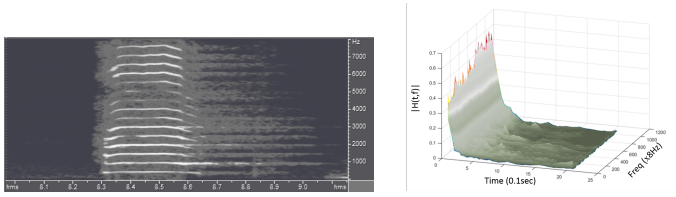

Fig. 6. Left: Spectrogram of the word "help" shouted in a room with a glass wall 6 feet behind the speaker. Reverberation causes significant spectral smearing, seen towards the right of the harmonics. Right: Room impulse response estimated through de-convolutive factorization of a signal.

Deducing the specifications of the speaker's equipment Voices may be degraded or modified in different ways by the devices used to transmit or record them. A measurement of voice quality, which comprises multiple aspects of voice [19] in the speech portions of the recording, and of the characteristic instrument signatures in the non-speech portions

of any recording, can yield important information about the instrumentation used to generate, transmit or record the voice.

V. Challenges that remain

None of the problems mentioned in this paper are fully solved. In the following, we enumerate some selected challenges that we have not discussed in this paper.

1) **Disambiguating the effects of all speaker parameters on voice**: Since a multitude of factors influence the voice signal, it is important to devise techniques to disambiguate their effects. The solutions would require the parameters to be *jointly* estimated, for which accurate algorithms must be designed.
2) **Noise compensation for profiling**: This includes dealing with poor signal-to-noise ratios, clipping, distortion, missing data or dropped data, compensation for reverberation effects etc. in a way that the profiling relevant information in the acoustic signal is not compromised.
3) **A better understanding of distressed speech**: There are subtle differences between real and simulated distress in human voice. Understanding these differences will lead to better techniques to disambiguate between an authentic call and a hoax call.
4) **Age correction**: Often, profiling must be done from voice samples that have been recorded significantly earlier in the past. Effective age-correcting mechanisms are required to deduce the *current* bio-relevant parameters of the speaker.
5) **Voice disguise**: Miscreants often disguise their voice as mentioned earlier, sometimes using commercially available devices that modify pitch and de-identify speech in other ways. The techniques for profiling must be made agnostic to these.
6) **Style recognition**: Each person has an a characteristic style of presentation, which generally remains the same when the individual is in different states, e.g. sober, intoxicated, or attempting disguise. In our observation, hoax calls are no exception. Ways to quantify style must be devised.
7) **Accent recognition**: Accent is a moving target. As people mingle geographically, their accents become diffuse and difficult to place. Traditional methods of accent recognition are no longer useful in this rapidly changing scenario. More realistic data needs to be obtained, and better analysis techniques need to be devised.
8) **Characterization of mental states induced by near-term factors**: Techniques must be devised for the characterization of mental states that are specifically relevant to profiling hoax callers. Examples include mental states induced by intoxication, medications or recreational drugs, emotions such as anger and behavioral patterns such as lying, evasion, deception, malice etc.

A. Legal issues and mindfulness thereof

Hoax calls are federal crimes in USA. The criteria of acceptability of our results must therefore be scientific *and* legal. The latter are outlined by the (controversial) Daubert and Fyre standards [20], [21] (depending on the state/region) that define the acceptability of scientific analysis. According to the Dec 1, 2011 amendment to the Daubert criteria, to be acceptable in a court of law, scientific results must: a) be based on sufficient facts or data, b) be a product of reliable principles and methods, and c) the methods must be reliably applied to the facts of the case. In compliance, we ensure that our techniques are accompanied with quantitative measures of confidence.

Acknowledgment

Dr. Singh was supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI002-07, via the Command, Control and Interoperability Center for Advanced Data Analysis (CCICADA).

References

[1] R. F. Coleman, J. H. Mabis, and J. K. Hinson, "Fundamental frequency-sound pressure level profiles of adult male and female voices," *Journal of Speech, Language, and Hearing Research*, vol. 20, no. 2, pp. 197–204, 1977.
[2] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: use of the long-term average spectrum (ltas)," *Journal of Voice*, vol. 10, no. 1, pp. 59–66, 1996.
[3] Telecommunications Industry Association and others, *Land mobile FM or PM communications equipment measurement and performance standards*. Telecommunications Industry Association, 2010.
[4] P. Delattre, "Coarticulation and the locus theory," *Studia Linguistica*, vol. 23, no. 1, pp. 1–26, 1969.
[5] R. Singh, B. Raj, and D. Gencaga, "Forensic anthropometry from voice: an articulatory-phonetic approach," in *Biometrics & Forensics & De-identification and Privacy Protection, MIPRO2016*. Croatia, 2016.
[6] J. Mercer, "Popeye the Sailor Man," *Voice art rendition*, 1935.
[7] Linguistic Data Consortium, "TIDIGITS," https://catalog.ldc.upenn.edu/LDC93S10, 1993.
[8] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *4th IEEE International Workshop on Biometrics and Forensics (IWBF)*, Cyprus, March 2016.
[9] M. Sonderegger and J. Keshet, "Automatic discriminative measurement of voice onset time." in *INTERSPEECH*, 2010, pp. 2242–2245.
[10] Linguistic Data Consortium, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S1, 1993.
[11] A. Kumar and B. Raj, "A novel ranking method for multiple classifier systems," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1931–1935.
[12] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. West Sussex, UK: John Wiley & Sons, 2012.
[13] B. Raj, T. Virtanen, and R. Singh, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Proc. Interspeech*, 2011.
[14] S. Chaudhuri, R. Singh, and B. Raj, "Block-sparse basis sets for improved audio content estimation," in *ICASSP*, 2013.
[15] S. Chaudhuri, *Structured Models for Audio Content Analysis*. Carnegie Mellon University: PhD dissertation, 2011.
[16] J.-C. Junqua, "The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication*, vol. 20, no. 1, pp. 13–22, 1996.
[17] M. Long, *Architectural acoustics*. Elsevier, 2005.
[18] K. Kumar, B. Raj, R. Singh, and R. M. Stern, "An iterative least-squares technique for dereverberation," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, pp. 5488–5491.
[19] E. H. Buder, "Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990," *Voice quality measurement*, pp. 119–244, 2000.
[20] S. Jasanoff, "Law's knowledge: science for justice in legal settings," *American journal of public health*, vol. 95, no. S1, pp. S49–S58, 2005.
[21] K. Pyrek, *Forensic science under siege: The challenges of forensic laboratories and the medico-legal investigation system*. Academic Press, 2010.