Speech Communication 101 (2018) 94–108

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks☆

John H.L. Hansen*,[a], Hynek Bořil[a,b]

[a] Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering, University of Texas at Dallas, Richardson, TX, USA
[b] Pioneer Speech Signal Processing Laboratory (PSSPL), Electrical and Computer Engineering Department, University of Wisconsin–Platteville, USA

ARTICLE INFO

Keywords:
Realism
Speech recognition
Human-computer interaction
Computational paralinguistics

ABSTRACT

Recent years have witnessed notable advancements in the areas of speech, speaker and language/dialect recognition. However, many of the emerging scientific principles appear to be drifting to the sidelines with the assumption that access to larger amounts of data is all that is required to address a growing range of issues relating to new scenarios. This study surveys several challenging domains in formulating effective solutions in realistic speech data, and in particular the notion of using naturalistic data to better reflect the potential effectiveness of new algorithms. Our main focus is on intra-speaker mismatch and speech variability issues due to (i) differences in noisy speech with and without Lombard effect and a communication factor, (ii) realistic field data in noisy and increased cognitive load conditions, (iii) speech variability introduced by whispered speech, and (iv) dialect identification using found data. Finally, we study speaker–environment and speaker–speaker interactions in a newly established, fully naturalistic Prof-Life-Log corpus. The specific outcomes from this study include an analysis of the strengths and weaknesses of simulated vs. actual speech data collection for research.

1. Introduction

The field of speech and language processing technology has evolved significantly over the past several decades. The primary mode of voice capture is still maybe through voice communications via handheld smartphones, however voice enabled devices and systems continue to expand into homes, vehicles, workplaces, and public locations (i.e., kiosks, info stations, etc.). Also, video and voice capture via mobile technology continues to expand at a rapid pace, resulting in a diversity of audio data never seen or expected by speech and language processing technologies over the past two decades. As such, it is clear that the primary challenge in almost any speech, speaker or language processing and classification task is the ability to formulate a solution that overcomes mismatch between training and test conditions. Speech feature extraction, model training and development, and classification strategies have progressed significantly over the past fifty years, yet the overriding challenge continues to be the ability of speech and language algorithms to be *robust* as either speaker, technology (e.g., voice-capture), or environment based mismatch is introduced. Also, recent advancements in deep learning for speech recognition have increased the requirements for use of significantly more training data, requiring many developers to "lower the bar" on acceptable data in order to train current discriminating systems. Knowing where, how, and who contributed to the audio data plays a critical role in the resulting acoustic models, and therefore robustness of the ultimate solutions.

Why should speech researchers be concerned today? The primary reason is the overwhelming availability of *found* data in the field. When expanding in the 1980's through 1990's, speech research still focused on the formulation of carefully collected speech data in order to construct acoustic and language models for effective algorithm development. Today, there is an overwhelming and exponentially growing amount of speech data freely available, and a greater temptation to simply use whatever data is available to address a specific research task. The old expression, "you get what you pay for" holds true in this context, since found data typically has limited meta-data information. However, as this study will show, researchers need to exercise caution, since mismatch is ever present. Data resource consortia, such as LDC, take great care in collecting, transcribing and organizing speech and language data. However, if researchers use data for purposes other than what a specific corpus was originally collected for, they may in fact be

☆ This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. A preliminary short version of this article was published in the special session "Realism in Robust Speech Processing" in ISCA INTERSPEECH 2016 (Hansen and Bořil, 2016).

* Corresponding author.
E-mail address: John.Hansen@utdallas.edu (J.H.L. Hansen).
URL: http://www.utdallas.edu/~john.hansen (J.H.L. Hansen).

https://doi.org/10.1016/j.specom.2018.05.004
Received 13 October 2017; Received in revised form 17 May 2018; Accepted 24 May 2018
Available online 29 May 2018

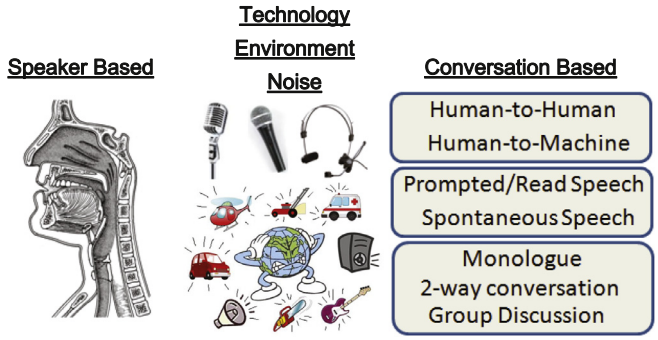

Fig. 1. Mismatch in speech and language processing: (i) speaker, (ii) technology, environment, noise, (iii) conversation-based.

constructing an irrelevant solution (e.g., Bořil et al., 2012).

In Fig. 1, we highlight a range of factors that can either individually or in combination contribute to audio stream mismatch between train and test. These can be partitioned into three broad classes: (i) speaker based, (ii) conversation based, and (iii) technology, environment or noise based. The speaker recognition community has used "intrinsic" mismatch to reflect speaker based changes, and "extrinsic" to represent either technology (microphone, recording/data capture, transmission, etc.) or environment and room factors. *Speaker-based variability* (SV) (see Fig. 2) is caused by (a) differences across speakers—*inter-speaker variability*, and (b) within-speaker variability—*intra-speaker variability*, representing the range of changes in how each individual speaker produces their speech. Inter-speaker variability can be categorized into *personal variations* and *sociolinguistic variations* (Umesh, 2011). Personal variations are attributed to physiological differences between subjects in terms of their speech production systems (e.g., size of the vocal tract and larynx (Adank et al., 2004)) and auditory systems (e.g., normal-hearing subjects versus cochlear implant users (Ruff et al., 2017)). Sociolinguistic variations, for example dialect (Hirayama et al., 2015) and accent (Najafian et al., 2014), are affected by such factors as regional and educational background or gender of the subject (Umesh, 2011). Another factor that contributes to both personal and sociolinguistic variations is the age of the subject (Pellegrini et al., 2012; Wagner, 2012; Bořil et al., 2011; Volín et al., 2017).

The major sources of intra-speaker variability include the following:

- *Stress—situational task or cognitive*—the subject is performing some task while speaking, such as operating a vehicle; hands-free voice input which can include cognitive (Bořil et al., 2010) as well as physical task stress (Hansen, 1996). Some forms of emotion, especially if they are a mixture, are also typically included in the domain of "stress" (Zhou et al., 2001; Womack and Hansen, 1999; Bou-Ghazale and Hansen, 2000; Hansen et al., 2012).
- *Vocal effort/style*—the subject alters their speech production from normal phonation in a deliberate controlled manner. This results in a vocal effort speaking style ranging from *whisper* (Fan and Hansen, 2013; Zhang and Hansen, 2011; Ghaffarzadegan et al., 2014; 2015) through shouted speech. Training data is typically neutral, so naturalistic or field data here could contain *whisper*, soft, loud, or shouted speech styles (Baghel et al., 2017; Hanilci et al., 2013). Each speaking style has a significantly different speech production configuration.
- *Lombard effect*—occurs when speech is produced in the presence of noise, and can be related to changes in vocal effort but because this is often a subconscious change in the production space due to the environment, it is listed separately (Hansen and Varadarajan, 2009; Bořil, 2008; Bořil and Hansen, 2010). It has also been known that speech produced in noise results in Lombard effect, but only recently has it been shown that both noise type and noise level result in different "flavors" of Lombard effect, which must be addressed for speech systems (Hansen and Varadarajan, 2009).
- *Speaking vs. singing style*—it has also been shown that speech production variability exists between speaking a specific text sequence versus singing that same text sequence (Mehrabani and Hansen, 2013). This impacts both speaker recognition as well as langauge ID systems.
- *Non-speech sounds*—studies have also explored speaker recognition solutions when the audio stream contains non-speech vocalizations such as screams (Hansen et al., 2017), whistles (Nandwana et al., 2015), etc. Non-speech sounds represent vocalizations which do not contain phoneme/text content, and could include coughs, whistles, screams, lip smacks, or other sounds humans make which are not speech.
- *Emotion*—the subject is communicating their emotional state while speaking (e.g., anger, sadness, happiness, etc.) (Hansen et al., 2000).
- *Physiology*—the subject has some illness (Lee et al., 2016), or is intoxicated (Zhang et al., 2017) or under the influence of medication; can include aging as well (Frederic Aman and Portet, 2013; Kelly and Hansen, 2016; Volín et al., 2017).

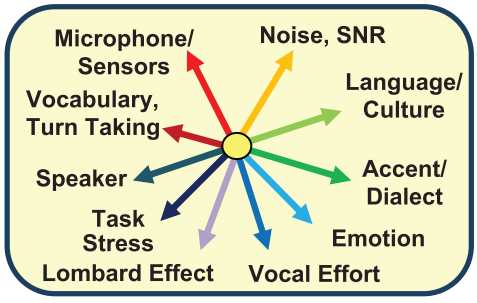

Fig. 2. Sources of speaker-based variability.

Conversation-based variability reflects different scenarios with respect to the voice interaction with either another person or technology, or differences with respect to the specific language, dialect, or accent spoken (Sulyman et al., 2014), and can include:

- *Human-to-human conversation*: two or more individuals interact (Shokouhi and Hansen, 2017); or one person speaks while addressing an audience. This scenario is affected by personal and sociolinguistic characteristics of the communicating parties (see inter-speaker variability discussed earlier in this section) such as their social status, gender, age, speech and hearing impairment, language or dialect spoken, if speech is read/prompted (through visual display or through headphones), spontaneous, conversational, monologue, 2-way conversation, public speech, group discussion. Another factor is the mode of the communication—a face-to-face communication, technology-mediated communication (Bordia, 1997) involving audio-visual or only audio channels (Banks et al., 2015).
- *Human-to-machine*: the subject is directing their speech towards a piece of technology (e.g., a spoken dialog system via a cell, smartphone, landline telephone, computer). This can include prompted speech—voice input to a computer; or voice input for telephone/dialog system (Bořil et al., 2010).

Conversational based issues represent additional challenges since the level of coarticulation will change depending on prompted vs. spontaneous speech, as well as confrontational style speech such as debates, etc. where there is an adversarial status between the speaking participants.

Technology- or external-based variability: includes how and where the audio is captured and range the following issues. *Electromechanical*—transmission channel, handset (cell, cordless, landline), microphone (Bořil et al., 2012; Kenny et al., 2007; Auckenthaler et al., 2000). *Environmental*—background noise (Rose et al., 1994; Liu

and Hansen, 2014) (stationary, impulsive, time-varying, etc.), room acoustics (Jin et al., 2007), reverberation (Greenberg et al., 2010; Sadjadi et al., 2012), distant microphone (Mirsamadi and Hansen, 2016). *Data quality*—duration, sampling rate, recording quality, audio codec/compression (Bořil and Fousek, 2006; Bořil, 2005).

Technology, environment, and noise factors do not only affect the process of sensing and transferring of the speech signal but often also impact the communicating parties. For example, presence of strong environmental noise will not only corrupt the sensed speech signal but also induce Lombard effect; delayed auditory feedback due to cochlear implant processing (Stone and Moore, 2002), strong room reverberation or a delayed feedback produced by a communication technology may negatively impact or even completely impair the speaker's ability to communicate, etc. For this reason, speaker-, conversation-, and technology, environment, and noise-based factors and their interactions need to be studied together to gain complete understanding of the speech signal variability and its potential mismatch with speech engine models.

Given the range of speaker, environment, acoustic, and technology based mismatch, what impact do these issues introduce to speech, speaker, and language recognition systems, and what steps can researchers do to minimize these issues? The remainder of this paper is organized as follows. Sections 2 and 3 study intra-speaker variability introduced by (i) environmental noise, (ii) communication scenarios, and (iii) emotions in terms of realistic speech acquisition as well as their impact on speech technology. Several examples of data acquisition protocols that to a certain extent depart from realistic conditions are given. Section 4 presents a case study on how poorly controlled or completely disregarded channel variability during data acquisition may result in a flawed experimental setup. Section 5 reflects on the number of issues discussed in the previous sections and presents a case study on a newly acquired naturalistic speech corpus that captures a range of real-world speech signal variabilities. Finally, Section 6 concludes the paper. The text presents a sequence of examples that demonstrate effects of speech signal variability on speech technology, and highlights potential departures from realistic conditions seen in some of the recent database designs.

2. Communication in noise

Recent years have witnessed a massive invasion of portable smart devices into our daily lives. The ever increasing computational power and broadband connectivity of these indispensable assistants makes them an ideal platform for hosting speech-enabled applications and services. While their portability is attractive and convenient to the users, it creates great challenges to the designers of the speech engines due to the enormous variety of conditions captured in the processed speech as the user transitions between different environments and communication scenarios.

To reflect the needs of the emerging market for speech technologies, researchers have been greatly focused on the development of speech processing techniques that would attain reasonable performance in adverse real world scenarios. Some of the major factors impacting efficiency of speech engines are speaker and channel variability, room reverberation (Kumar et al., 2011; Sadjadi et al., 2012), and environmental noise. While all these factors clearly impact engines running on portable devices, varying environmental noise may be often the most disruptive and hard to deal with element in the process (Li et al., 2014).

A design of efficient noise modeling and suppression techniques demands availability of a rich noisy speech data. The most convenient and economic way to acquiring noisy material is to mix clean speech recordings with noise samples. This provides an excellent control in terms of the desired signal-to-noise ratios and allows for reusing the same, easily accessible clean speech corpus for producing vast number of noisy mixtures with different noise samples without requiring the human subjects to re-record their utterances over and over. Since the speech component stays intact and only the additive noise is varied, this approach allows for studying isolated effects of noise separately from other types of variations that would naturally occur if the speakers had to reproduce their utterances. This approach to the creation of noisy speech samples has been widely popular. Some of the most prominent examples can be found in the speech recognition-oriented Aurora datasets. Aurora 2 was created by artificially contaminating clean TIDigits (Pearce et al., 2000); Aurora 4 followed the same approach on Wall Street Journal recordings (Parihar et al., 2004). Aurora 5 returned back to TIDigits and added simulated distortion factors including a hands-free microphone channel, transmission through a GSM channel, and room reverberation (Hirsch and Finster, 2005). Others have followed this trend to create challenging corpora for other application domains, such as those seen in the NIST Speaker Recognition campaigns (Hasan et al., 2013b). The contribution of the Aurora and NIST suites is undisputed and quite remarkable—they have provided unified development and evaluation frameworks for fair and transparent comparison of speech systems and significantly accelerated the advancement of speech technology. This being said, and as will be discussed in the following text, artificially mixing clean speech recordings with noise leaves out other factors that may be equally detrimental to the system performance. Solely relying on such data in system evaluations is likely to provide unrealistically optimistic results compared to real world adverse recordings and using such data in system design will lead to suboptimal performance in real world conditions.

2.1. Adding noise versus talking in noise

Clean speech recordings artificially contaminated with noise samples may provide a reasonable approximation to actual speech distortion by additive environmental noise, however, they will not capture the effects of noise on speech production. When speaking in noisy environments, speakers continuously adjust their speech production to maintain intelligible communication (Lombard effect (Junqua, 1993; Hansen, 1996)). Lombard effect has a prominent impact on a number of speech production parameters (Garnier, 2007; Lu and Cooke, 2008; 2009; Bořil, 2008).

During his initial experiments, Etienne Lombard noted that speakers' vocal changes in response to noise seemed to be unconscious (Lombard, 1911). This originated a theory that speech production is an automatic servomechanism controlled by auditory feedback. This seems to be supported by Pick et al. (1989) where speakers were unable to follow instructions to maintain constant vocal intensity across alternating periods of quiet and noise. On the other hand, Lane and Tranel (1971) observed significant differences in speech production for speakers who were communicating or just reading texts, suggesting that the reaction to noise is not purely automatic but rather consciously driven by the speakers' effort to maintain intelligible communication. The same study hypothesizes that the response to noise may be initially learned through the public loop (speaker–listener) and later becomes a highly practiced reaction. In Junqua et al. (1998), speakers were exposed to noise while communicating with a voice-controlled dialing system. The system utilized neutral-speech trained acoustic models and hence, would perform best when encountering matching neutral speech modality. The subjects were able to consciously compensate for the Lombard effect and lower their voices, in spite of the present noise, to reach efficient response from the system. This confirms the hypothesis in Lane and Tranel (1971) that speech production changes in noise are at least to some extent driven by a conscious response to the public loop. These observations lead to the definition of Lombard effect as stated in the previous paragraph and used in Junqua (1993); Womack and Hansen (1999). However, it remains unclear to what extent or in which conditions the speakers' speech production changes made in order to maintain intelligible communication are conscious of subconscious.

As observed already in Lane and Tranel's seminal paper from 1971 (Lane and Tranel, 1971), the rate of speech production variations under Lombard effect does not depend only on the type and level of environmental noise but also on a number of other factors such as the communication scenario, the way the noise is presented to the subjects, and the means of establishing auditory feedback. Moreover, Lombard effect is strongly speaker-dependent. A combination of all these factors cause a large variation, or sometimes even contradictions, in observed trends across Lombard effect studies. Taking this into account, the following paragraphs summarize the aspects of speech variability under Lombard effect that are agreed upon by a large body of literature, while the reader is being cautioned that studies presenting alternative observations to some of these trends may be found.

Speakers increase their vocal effort (Lombard, 1911; Dreher and O'Neill, 1957; Webster and Klumpp, 1962) and the increase is non-uniform across phone classes with vowels being typically more emphasized that consonants (Hansen, 1988; Junqua, 1993). The higher vocal effort is accompanied by increases in the fundamental frequency (Lombard, 1911). This is in part caused by the physiological relationship between the fundamental frequency of the glottal waveform and sub-glottal pressure and tension in the laryngeal musculature, which are elevated during higher vocal efforts (Schulman, 1985). In some studies, fundamental frequency of speech was reported to change almost linearly with vocal intensity (Gramming et al., 1987).

Lombard effect impacts temporal profiles of glottal waveforms (Cummings and Clements, 1990). Spectral energy typically migrates to higher frequencies, causing upward shift of the spectral center of gravity (Junqua, 1993; Lu and Cooke, 2008) perceptually related to the "brightness" of speech. This goes hand in hand with flattening of the spectral slope (Hansen, 1988; Pisoni et al., 1985; Summers et al., 1988). The first formant F_1 migrates upwards in frequency in Lombard speech (Schulman, 1985; Bond and Moore, 1990), and the rate of the shift is phone dependent (Junqua and Anglade, 1990; Bořil, 2008). The second formant F_2 may shift in either direction in frequency depending on the phonetic classes and other factors (Bořil, 2008; Junqua, 1993; Bond et al., 1989; Pisoni et al., 1985; Hansen, 1988; Hansen and Bria, 1990; Takizawa and Hamada, 1990). Bandwidths of the first four formants are reduced compared to neutral speech for most phones (Hansen, 1988; Hansen and Bria, 1990; Junqua, 1993; Bořil, 2008). Syllable and word durations are typically prolonged in Lombard speech (Dreher and O'Neill, 1957; Junqua and Anglade, 1990; Lane and Tranel, 1971; Hansen, 1988; 1996; Bořil, 2008).

Even if the environmental noise which triggered Lombard effect is suppressed in the recording or completely excluded (e.g., when the noise is produced to the subjects via closed-air headphones), the speech variability caused by Lombard effect may result in a severe mismatch with the neutral speech-trained acoustic models and cause the speech system to break (Bořil and Hansen, 2010).

Fig. 3 shows performance of a neutral-trained automatic speech recognition (ASR) system tested on TIMIT-like (Zue et al., 1990) utterances produced by speakers who were exposed to three levels of a highway (HWY), large crowd (CRD/LCR), and pink noise (PNK) played back through headphones (70, 80, and 90 dB SPL for HWY and CRD; 65, 75, 85 dB SPL for PNK). The experiment was conducted on a close-talk microphone channel with a high signal-to-noise ratio (SNR) and involved speech recordings from 31 native speakers of American English (25 females, 6 males) as captured in the UT-Scope Lombard Effect set (Hansen and Varadarajan, 2009) (see Bořil and Hansen, 2011 for more details on the ASR experiment). The word error rate (WER) can be seen to grow rapidly from the baseline clean *Neutral* condition once the speakers are exposed to increasing noise levels. Note that in all noisy conditions utilized in this experiment, the speech signal retains a high SNR as the rate with which the subjects increase their vocal level masks the already minimal cross-talk between the closed-air headphones and the close-talk microphone.

Similar impacts of Lombard effect have been observed in speaker

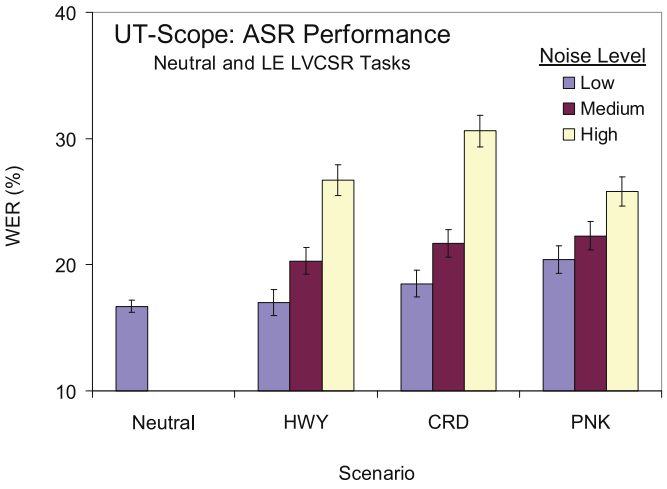

Fig. 3. Talking in noise: ASR performance on clean neutral and clean Lombard speech UT-Scope tasks; a TIMIT language model; 95% confidence intervals.

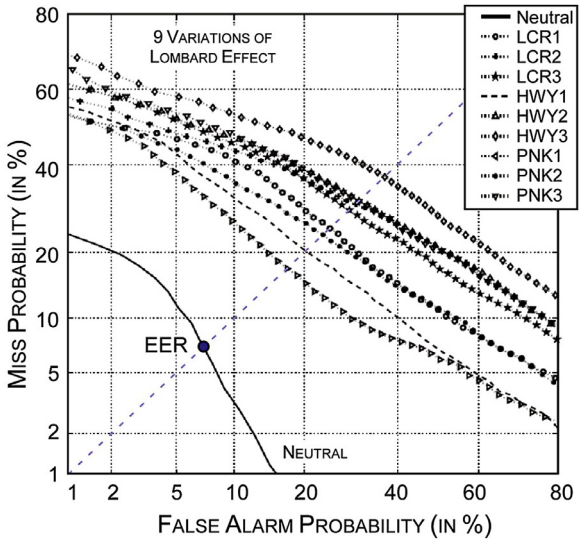

Fig. 4. Talking in noise: SV performance on clean neutral and clean Lombard speech UT-Scope tasks; 12 s test utterances; *Neutral*—clean neutral samples; *LCR1–3*–large crowd noise presented at 70, 80, and 90 dB SPL; HWY1–3–highway noise presented at 70, 80, and 90 dB SPL; *PNK1–3*—pink noise presented at 65, 75, 85 dB SPL (speakers exposed to these noise levels using open-air headphones; all recordings are noise-free).

verification (SV). Fig. 4 presents detection error tradeoff (DET) curves for a SV task on the UT-Scope database (Hansen and Varadarajan, 2009). This experiment involved 30 subjects (19 males and 11 females) and similarly as in the ASR task discussed above, the speech variability due to Lombard effect results in degraded SV performance as a result of the increased mismatch between the classified speech samples and the reference speaker models and universal background model (UBM) (see Hansen and Varadarajan, 2009 for more details). It is safe to state that none of these noise-induced speech variations could be found in the Aurora or NIST datasets where the noise was artificially added to neutral speech recordings and hence, it is unclear what the actual performance of the systems tuned and evaluated on these sets would be in real world noisy conditions.

2.2. Talking in noise: reading versus communicating

Going one step further from simply mixing noise with clean speech, some studies record speech in realistic or simulated noisy conditions to

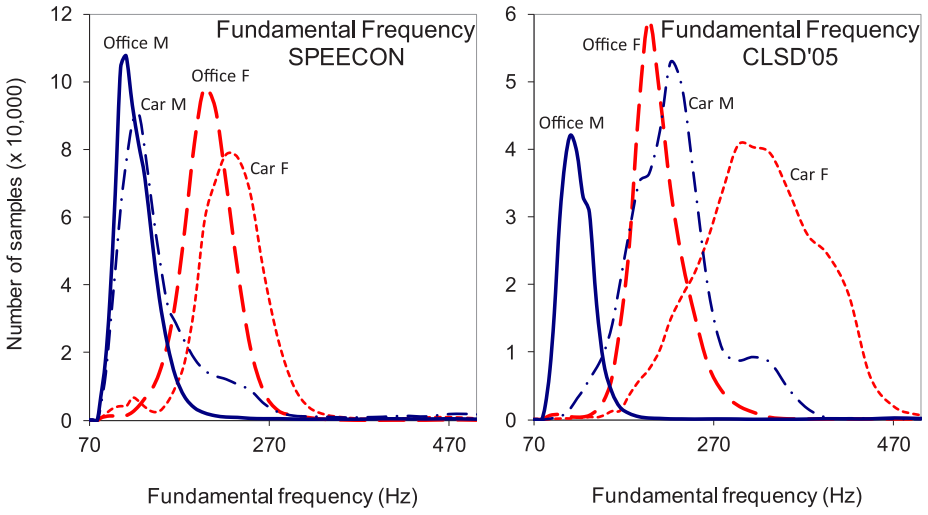

Fig. 5. Talking in noise: fundamental frequency in scenarios without (SPEECON) and with (CLSD'05) communication loop; *F/M*—female/male subjects.

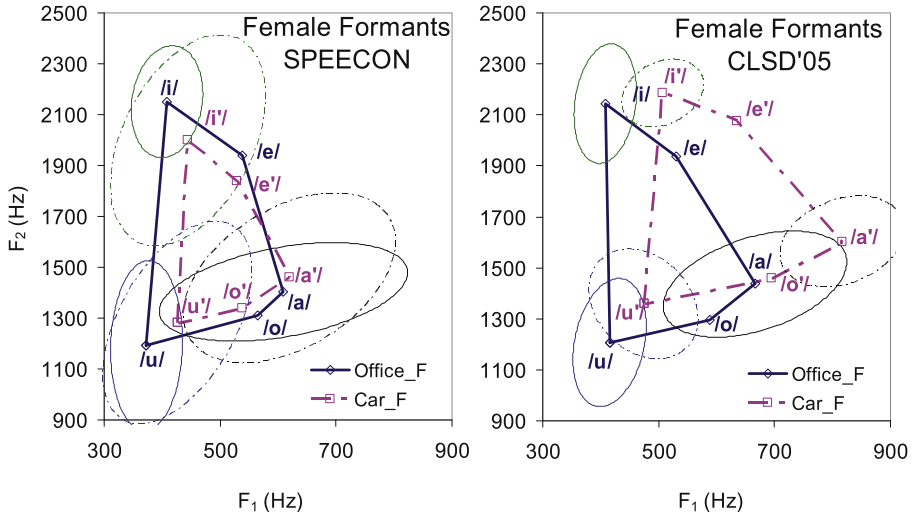

Fig. 6. Talking in noise: vowel locations in F_1–F_2 plane in female utterances in scenarios without (SPEECON) and with (CLSD'05) communication factor; *F/M*—female/male subjects; 1-σ ellipses estimated to cover 39.4% of samples.

capture noise-induced speech production variability. Yet, adopting the concept of many neutral speech databases, the recorded subjects are asked to read prompts in noise without being engaged in an actual communication (Junqua, 1993; Hansen, 1996; Lu and Cooke, 2008; Barker et al., 2015; 2017). While the subjects will respond to the background noise to some extent, especially when instructed to imagine they are addressing some person or a piece of technology (Barker et al., 2017), they are reading the prompts without any feedback that would help them adjust their voices in a way to effectively convey the linguistic content to others over the noise. Since this scenario lacks the communication loop, the speech production adjustments are left solely to the judgment of the speaker. As shown in Lane and Tranel (1971); Bořil (2008) and when comparing for example Dreher and O'Neill (1957) and Webster and Klumpp (1962), presence or lack of communication will result in considerably different speech adjustments in response to noise. In the context of speech technologies, one can assume that the subjects mostly engage in communication with the device (human–machine interactions) or with other humans (e.g., processing radio or TV programs, forensic analysis of recorded dialogues). For this reason, it is instrumental to include communication loop in the acquisition of naturalistic speech corpora. An example of the different production adjustments for scenarios that involved or did not involve communication loop is shown in Figs. 5 and 6.

The figures study fundamental frequency (F_0) and vowel locations in the F_1–F_2 formant plane for utterances from Czech SPEECON (ELRA, 2008) and the Czech Lombard Speech Database (CLSD'05) (Bořil, 2008; Bořil and Pollák, 2005). In both datasets, the subjects produced utterances in (i) a clean office environment and (ii) when exposed to a car noise. The SPEECON recording protocol required the subjects to read the prompts to themselves and did not involve a communication loop. In CLSD'05, the subjects were required to communicate the prompts to a listener who was exposed to the same noise. The listener was instructed to ask for a repetition if the utterance was not intelligible to them. It is noted that the SPEECON and CLSD'05 datasets capture similar scenarios in terms of noise types, but they cannot be compared in absolute terms since in the SPEECON case, the recording was conducted in an actual car and in CLSD'05, pre-recorded car noise samples were produced to the subjects via a headset that contained also speech feedback. The purpose of presenting the SPEECON and CLSD'05 analyses side-by-side in Figs. 5 and 6 is to give the reader a notion on the rate of speech parameter changes in the case of speech read in clean versus noise conditions and when communicating in clean and noisy conditions. A study Cooke and Lu (2010) reveals similar trends for an experimental setup where the same noisy conditions and recording equipment were used in scenarios with and without communication and the communicative scenario yielded

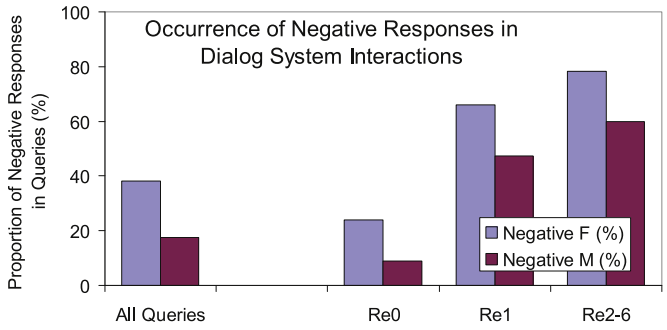

Fig. 7. Talking in noise: proportion of negative interactions with a dialog system as a function of gender; F/M—female/male subjects.

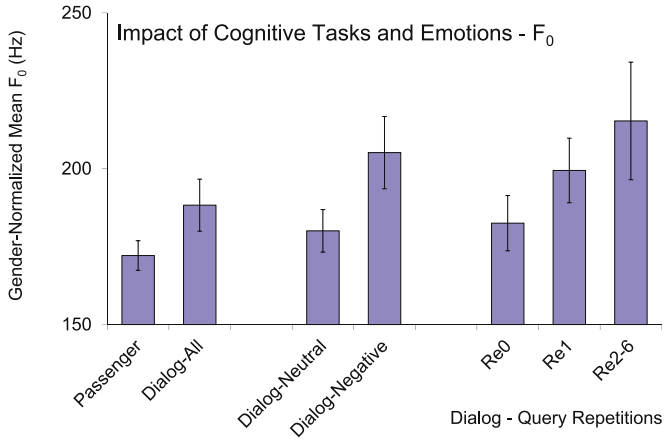

Fig. 8. Talking in noise: human-human versus human-dialog system communication (UT-Drive corpus); fundamental frequency; error bars represent 95% confidence intervals.

significantly higher rate of speech variation in Lombard speech than the non-communicative scenario. Similar conclusions can be found in Lane and Tranel (1971).

It can be observed in Figs. 5 and 6 that the F_0 changes and the formant shifts in the communication scenario (CLSD'05) are more pronounced and also more consistent, yielding relatively more compact 1-σ ellipses (see Bořil, 2008 for more details).

2.3. Talking in noise: communication scenarios

As discussed in the previous section, the presence or lack of the communication loop will affect the way we speak. Moreover, different communication scenarios may be associated with different vocal adjustment strategies. For example, speech addressed to infants (motherese) tends to have notably different characteristics compared to adult-directed speech (Narayan and McDermott, 2016). Communication parties may adapt their acoustic-phonetic spaces towards each other via phonetic convergence (Pardo, 2013). Their speech production will be also impacted by their mutual relationship and differences in social status (Leongomez et al., 2017).

Figs. 7, 9, 10,11 present results of speech analyses on 68 subjects from the UTDrive database (Angkititrakul et al., 2007). The subjects were driving in real traffic while performing various secondary tasks (i.e., tasks conducted in addition to driving the vehicle). Fig. 7 summarizes the proportion of conversational turns between the drivers and an automated dialog system that were perceived by a human transcriber as negative. The drivers were instructed to call a commercial automated dialog system and retrieve a specific information. Due to the occasional ASR errors, the drivers had to repeat some of the queries until they succeeded in completing the task. Such repetitions sometimes

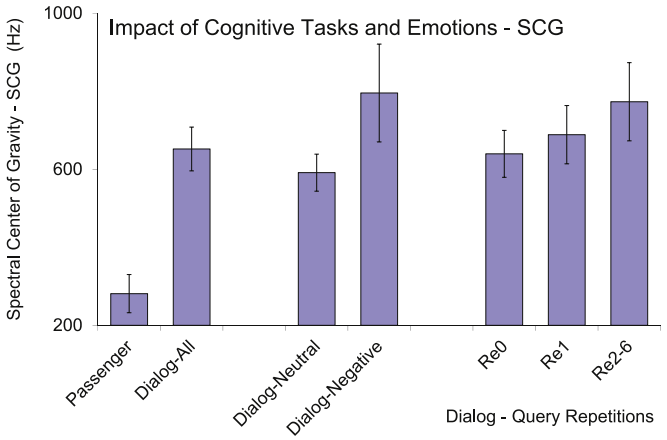

Fig. 9. Talking in noise: human-human versus human-dialog system communication; spectral center of gravity (SCG); error bars represent 95% confidence intervals.

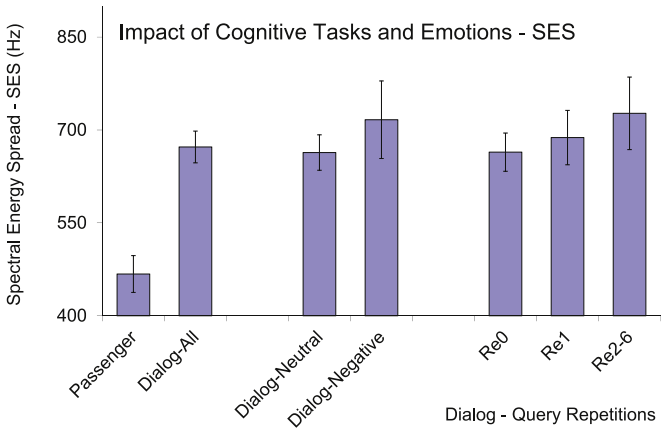

Fig. 10. Talking in noise: human-human versus human-dialog system communication; spectral energy spread (SES); error bars represent 95% confidence intervals.

resulted in notable driver frustration. Not surprisingly, the frustration was observed to further build up with the number of involuntary repetitions. The figure breaks down the human–dialog system interactions by gender. It can be seen that across all conditions (no repetition, 1–6 query repetitions needed), there were notable differences between the genders in terms of the tendency to get frustrated with the dialog system.

Figs. 8–11 study fundamental frequency, spectral center of gravity, spectral energy spread, and mean duration of voiced segments estimated for the driver's speech while (i) casually talking to the passenger (*Passenger*), and (ii) calling the dialog system (*Dialog*).

These figures confirm that the communication mode (talking to a passenger versus a dialog system), emotions (neutral/negative), and the number of query repetitions, no repetition (*Re0*), one repetition (*Re1*) and 2–6 repetitions (*Re2–6*), all affect speech production (see Bořil et al., 2010 for more details).

3. Whispered speech

Similar to speech under Lombard effect, whispered speech represents yet another mode of speech with altered vocal effort. While Lombard effect will be encountered in communication in noisy environments, *whisper* is often used in relatively quiet environments to share discreet or private information. *Whisper* is often used in human–human communication but may be equally attractive as a means for human-machine interactions, especially when using handheld

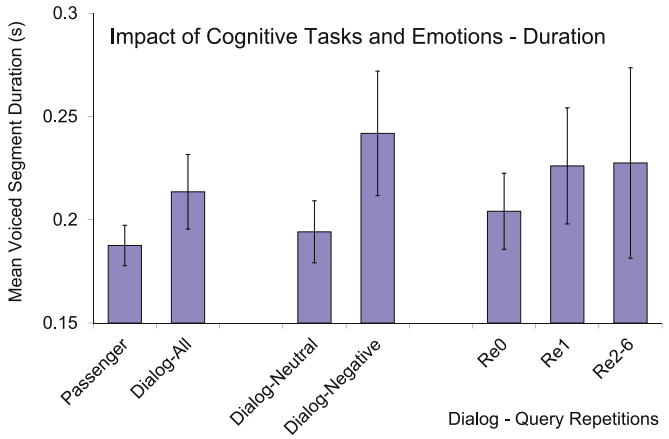

Fig. 11. Talking in noise: human-human versus human-dialog system communication; mean voiced segment durations; error bars represent 95% confidence intervals.

devices such as smartphones in open-office settings, company meetings, or quiet places such as libraries.

The majority of current speech engines are oriented towards neutral speech and perform poorly when exposed to *whisper*. Neutral (modal) speech is dominated by voiced sounds where the airflow from the speaker's lungs causes vibration of the vocal folds. These vibrations are the excitation source of the vocal tract. On the other hand, in pure *whisper* (i.e., speech lacking any voiced components), the glottis is constantly open and the turbulent flow of the passing air provides excitation for the articulators (Morris and Clements, 2002). In addition to missing the periodic excitation, *whisper* is characterized by different prosodic cues (Heeren and Lorenzi, 2014), phone durations (Lee et al., 2014), distribution of energy across phonetic classes, spectral tilt, and formant center frequencies and bandwidths (Zhang et al., 2010; Fan and Hansen, 2010; 2011; Ito et al., 2001; Eklund and Traunmuller, 1997; Ito et al., 2005; Lim, 2011; Matsuda and Kasuya, 1999; Morris and Clements, 2002), as well as altered distribution of phones in the F_1–F_2 formant space (Sharifzadeh et al., 2012).

Fig. 12 shows an example of neutral and whispered speech waveforms capturing the same linguistic content. The accompanying spectrograms demonstrate how in the voiced regions of the neutral utterance, a dominant portion of the spectral energy is concentrated at lower frequencies (mostly below 2 kHz, with visible harmonics of the fundamental frequency), while in the whispered utterance, the spectral energy is distributed across a broader range of frequencies and lacks any harmonic

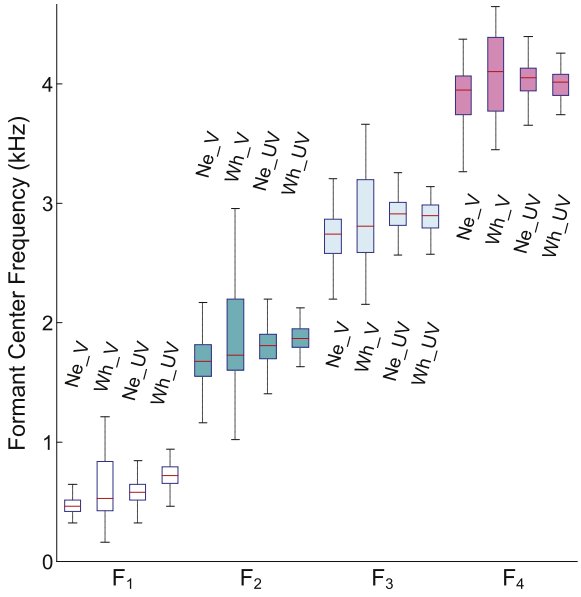

Fig. 13. Formant center frequency distributions; *Ne/Wh*—neutral/*whisper; V/UV*—voiced/unvoiced.

structure. It can be also seen that the unvoiced /z/ (from "dishes") appearing at the end of both the neutral and whispered utterances has a similar structure in both, suggesting that in spite of the likely effects of co-articulation with the adjacent phones, unvoiced consonants in neutral and whispered speech may share similar characteristics.

Speech engines trained for neutral speech perform poorly on *whisper* due to the severe acoustic mismatch between the neutral acoustic models and processed whispered samples. Studies on whispered speech recognition mostly utilize model adaptation (Ito et al., 2001; 2005; Lim, 2011; Mathur et al., 2012; Jou et al., 2005), vector Taylor series transformations of the *whisper* samples towards neutral (Yang et al., 2012), or inverse filtering (Galic et al., 2014) and (Grozdic et al., 2014). Hybrid deep neural network–hidden Markov models (DNN–HMM) were recently explored for *whisper* ASR in Lee et al. (2014); Ghaffarzadegan et al. (2017), and an audiovisual approach to speech recognition was taken in Tao and Busso (2014). Whispered speech processing has been studied also in the context of speaker identification (Fan and Hansen, 2010; 2011; 2013), automatic *whisper* island detection (Zhang et al., 2010), and modal speech synthesis from *whisper* (Morris and Clements, 2002).

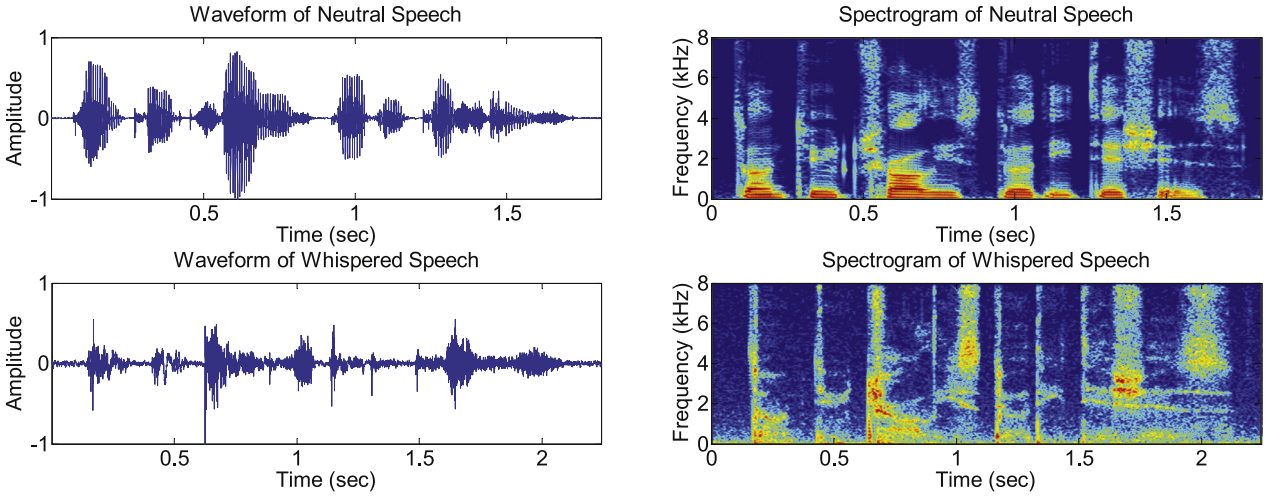

Fig. 12. Time domain waveform and spectrogram of neutral and whispered utterance *"Don't do Charlie's dirty dishes"* produced by one speaker.

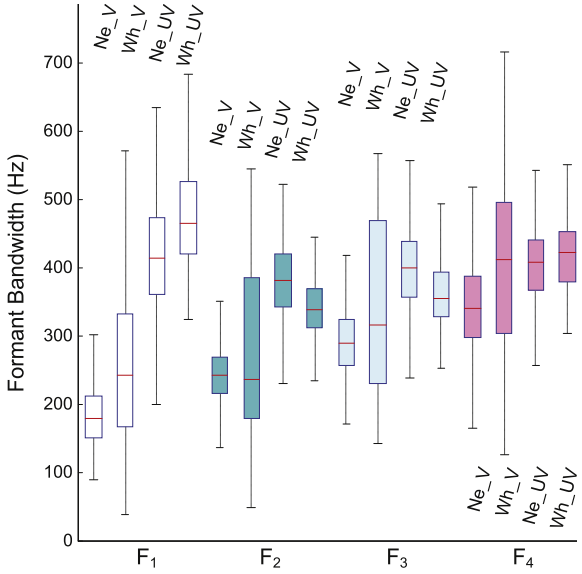

Fig. 14. Formant bandwidth distributions; *Ne/Wh*—neutral/*whisper; V/UV*—voiced/unvoiced.

Figs. 13 and 14 show boxplots of formant center frequency and bandwidth distributions in neutral and whispered speech drawn from the UT-Vocal Effort II dataset (Ghaffarzadegan et al., 2016). It is noted that the *whisper* samples in UT-Vocal Effort II are dominated by unvoiced speech but contain also a small portion of occasional voiced segments and hence, the figure includes those as well (denoted *Wh_V*—voiced *whisper*). When switching from neutral speech to *whisper*, nearly all neutral voiced *Ne_V* segments in modal speech will be replaced by whispered unvoiced *Wh_UV* segments and it is easy to infer that on average, all the first four formants will be traveling upwards in frequency (Fig. 13) and their bandwidths will be, due to the noise-like nature of *whisper*, broadened (Fig. 14) as far as the neutral speech is dominated by voiced segments, which is typically the case. Comparison of neutral unvoiced *Ne_UV* segments and whispered unvoiced *Wh_UV* segments reveals that there is an prominent increase in F_1 and F_2 center frequencies, confirming the observations from Sharifzadeh et al. (2012), while F_3 and F_4 exhibit a slight shift downwards for *whisper*. The increase in F_1, F_2 follows a similar trend as seen in Lombard speech (Bořil and Hansen, 2010). These two modalities, while being quite different in a number of aspects (Lombard speech being louder and dominated by voiced segments compared to neutral speech while *whisper* being just the opposite), they may share similar strategies in terms of the jaw and tongue configurations during speech production. Note that the first formant center frequency varies inversely to the vertical position of the tongue and the second formant frequency increases with tongue advancement (Kent and Read, 2002; Sulyman et al., 2014; Hansen et al., 2015). Some older literature hypothesizes that the location of the third formant F_3 is less dependent on the linguistic content and more correlated with the vocal tract length in speakers (Claes et al., 1998; Eide and Gish, 1996). However, a recent study involving quantitative analyses of magnetic resonance imaging (MRI) does not seem to confirm this hypothesis (Hatano et al., 2012). The same study finds either negative or positive correlations between the MRI-measured vocal tract length and F_3 and F_4 shifts depending on the vowel uttered. In this sense, even if, hypothetically, whispering lead to a consistent adjustment of the speaker's vocal tract length with respect to neutral speech, the measured F_3 and F_4 center frequencies could be still traveling either upwards or downwards in frequency depending on the linguistic contents. For this reason, we attribute the slight decrease in the whispered F_3, F_4 means to the specific whispered phonetic balance captured by the database.

Fig. 15 presents a distribution of the first two cepstral coefficients c_0 and c_1 for neutral speech and *whisper*. Neutral speech combines voiced and unvoiced phone classes, which leads to mildly bimodal distributions of the two lowest cepstral coefficients. On the other hand, a pure *whisper* is dominated by unvoiced segments with one prominent mode per coefficient. This strong unimodality is likely causing a greater confusability between broad phonetic classes in *whisper* but also between whispered and silence segments (note that the silence and unvoiced speech distributions in the two bottom plots have similar contours and strongly overlap). The fundamental difference between the neutral and whispered cepstral distributions causes mismatch between *whisper* and neutral-trained acoustic models of speech, speaker, and language recognition systems and degrades their performance (Ghaffarzadegan et al., 2016; 2015; Fan and Hansen, 2011).

This concludes our survey of selected speech modalities in this section (neutral speech, Lombard effect speech, *whisper*, emotional speech, speech under cognitive load). Other talking styles can be found in literature. For example, the reader is encouraged to reach for (Zhang and Hansen, 2007) to learn more about *soft, loud*, and *shouted* speech in the context of speech technologies, (Hasan et al., 2013a) to study the effects of *arousal*, and (Hansen et al., 2012) to see the impact of a non-acted *stress* on speech production.

4. Channel characteristics

Channel variability is mostly perceived as a negative factor that increases mismatch in speech systems. Channel characteristics are affected by the room impulse response and the related reverberation effects (Sadjadi et al., 2012), transmission channel parameters and microphones used (Junqua, 2002), as well as the distance and orientation of the speaker with respect to the microphone or microphone array, most recently studied in the context of distant speech recognition (Ravanelli et al., 2017).

It is fairly standard for modern speech corpora to incorporate various channels for development and evaluations of robust speech engines, be it real channels used in the simultaneous capture of the speech material or simulated channels (e.g., ELRA, 2008; Hirsch and Finster, 2005; Hasan et al., 2013b; Barker et al., 2015). The speech community is well aware of the issues stemming from channel variability and a vast number of channel normalization and modeling techniques have been devised since the dawn of speech engineering. This being said, in some applications, channel characteristics may be leveraged in gathering valuable information about a particular environment or the recording equipment used during the data acquisition (e.g., "environmental sniffing" (Akbacak and Hansen, 2007)). On the other hand, there are situations where the speech corpus designers have good reasons to conceal any information about the origin of the individual samples. One example can be speaker, language, or dialect recognition evaluation campaigns. Here, the tested systems are expected to make decisions solely based on the speech contents of the samples while being, in an ideal case, immune to channel and noise characteristics present in the signal.

Fig. 16 demonstrates a real world example where channel characteristics "leaked" information about the dialect being spoken in the recordings. The left-hand side of the figure details long-term channel transfer functions estimated for LDC's conversational telephone speech (CTS) corpora capturing four Arabic dialects (Iraqi, Levantine, Gulf, Egyptian). As found in Bořil et al. (2012), processing just silence segments from these datasets is sufficient for a highly accurate dialect identification (DID) since the channel characteristics are perfectly correlated with the respective dialects. This is an example of using good data (the datasets were recorded in separate campaigns using different equipment and intended for ASR applications) in a wrong context (see examples in Biadsy et al., 2009; Biadsy et al., 2011; Akbacak et al., 2011). The right-hand side of the figure shows channel characteristics of an in-house Pan-Arabic corpus that was collected with the dialect

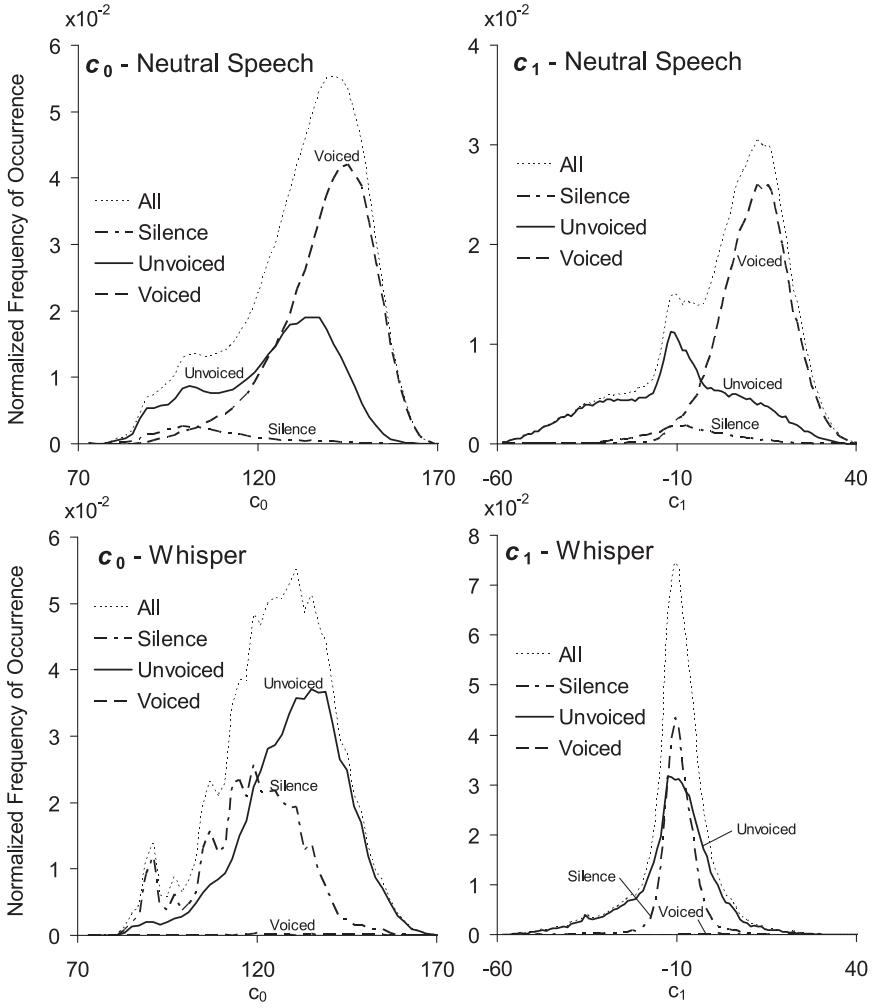

Fig. 15. Normalized cepstral distributions of broad acoustic classes in neutral and whispered speech.

identification task in mind and utilized a fixed recording setup for all scenarios. Here the channel characteristics do not reveal any useful information with respect to the dialect being spoken and the dialect identification systems need to focus solely on the speech content to make any decisions (Bořil et al., 2012).

5. Prof-Life-Log: a case study with naturalistic speech corpus

Up to this point, we have discussed several prominent sources of intra-speaker variability and related issues with realistically capturing such variability in speech corpora. We have studied the impact of intra-

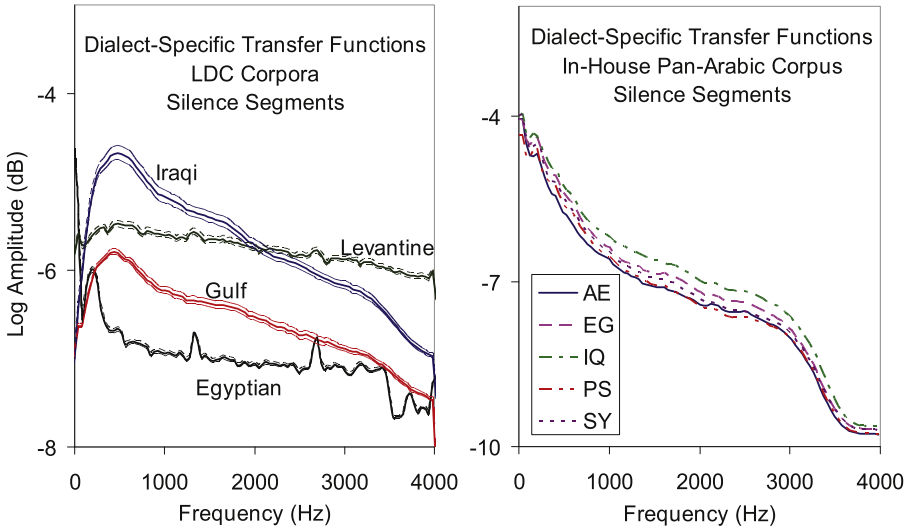

Fig. 16. (Left) Dialect-specific channel characteristics in Arabic CTS corpora—dashed lines are ± 5σ intervals; (right) channel characteristics in-house Pan-Arabic corpus capturing dialects of United Arab Emirates (AE), Egypt (EGY), Iraq (IRQ), Palestine (PS), and Syria (SY).

speaker variability on speech technology and have also given an example of how poorly planned or disregarded channel variability in speech data acquisition may lead to a flawed experimental framework. This section presents a case study on a recently collected, fully naturalistic corpus that addresses some of the issues discussed in the previous sections. Prof-Life-Log (Ziaei et al., 2013) captures audio recordings of entire work days of a university professor. The speech acquisition is conducted via the LENA (Language Environment Analysis) unit (Xu et al., 2012) fixed onto the subject's clothes. The device records the primary subject carrying the device, as well as those interacting with him (secondary speakers) together with all the surrounding sounds. The recorded interactions are fully natural and completely unscripted. The subject routinely interacts with his colleagues, students, and acquaintances in various environments and social settings.

The LENA unit captures speech using a single microphone channel. The analyses were conducted on recordings from four selected environments—*Office, Cafeteria, Walking,* and *Car* captured within one work day (12 h of audio). The environment information was perceptually labeled by expert human annotators. An automated system (Ziaei et al., 2013) was used to detect speech islands and perform primary vs. secondary speaker diarization. The operating point of the speech activity detector was set to keep a low rate of false alarms (~ 2.1%) to assure that the segments sent for diarization contain actual speech. The diarization error rate was estimated to be around 15% (based on evaluation on 6 h of manually annotated data). The fundamental frequency analyzed below was estimated for both primary and secondary speakers using Talkin's RAPT algorithm based on the Normalized Cross-Correlation Function (NCCF) (Talkin, 1995) and the formant tracking was performed by the Talkin's linear prediction and dynamic programming based formant tracking algorithm (Talkin, 1987), both implemented in WaveSurfer (Sjolander and Beskow, 2000). Due to the nature of the recording setup, the analyzed trends represent human-environment and human-human interactions in natural communication settings. It is expected that the speech production parameters estimated from the segmented recording session will be, to a certain extent, affected by the presence of the environmental noise, especially in the case of the secondary speakers that are further away from the microphone worn by the primary speaker. To reduce the impact of the analysis errors on the observed trends, means of the parameters extracted across all segments of the particular class of interest as captured in the 12-h session are presented and where deemed relevant, accompanied by their 95% confidence intervals or boxplots.

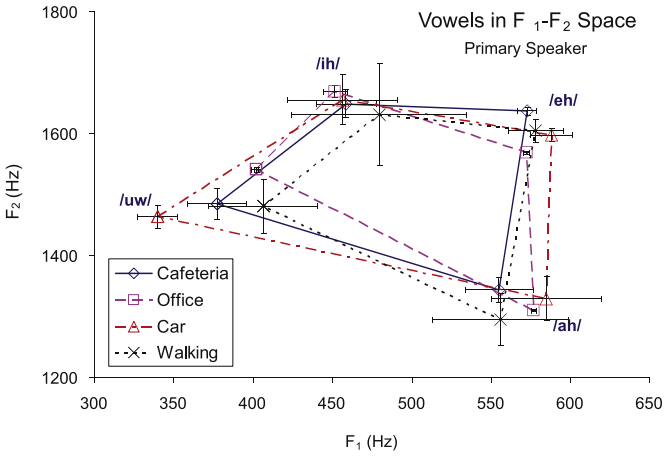

Fig. 18. Prof-Life-Log: environment-induced vowel shifts in the F_1–F_2 formant space.

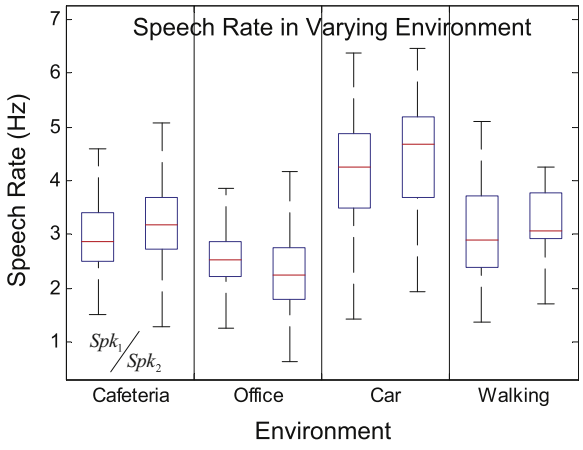

Fig. 19. Prof-Life-Log: speech rate (extracted from short-time energy envelopes, following (Heinrich and Schiel, 2011)) in primary and secondary speakers as a function of environment; boxplot edges—25th and 75th percentiles, central mark—median, whiskers—most extreme points that are not considered outliers.

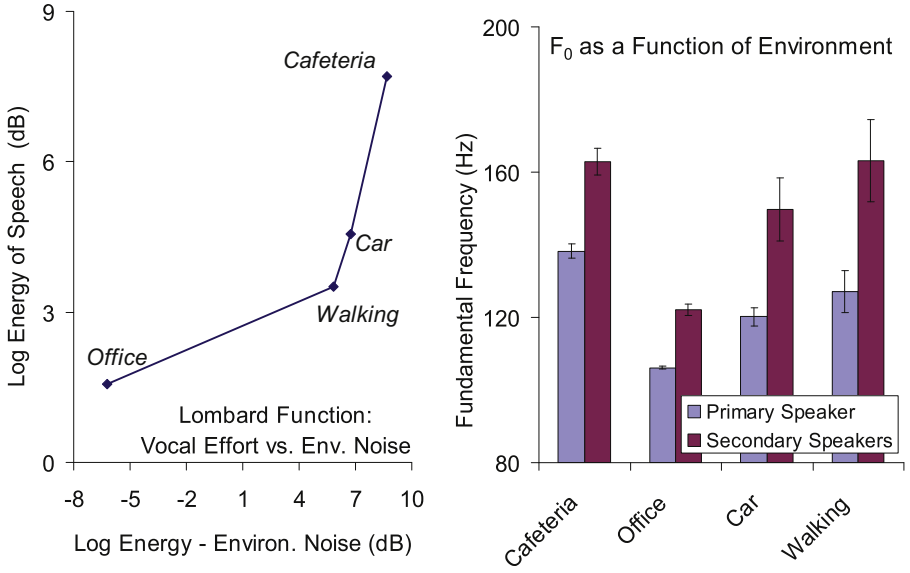

Fig. 17. Prof-Life-Log: (Left) Lombard function—vocal intensity as a function of environmental scenario; primary speaker; (right) fundamental frequency in primary speaker and secondary speakers in varying environments; error bars—95% confidence intervals.

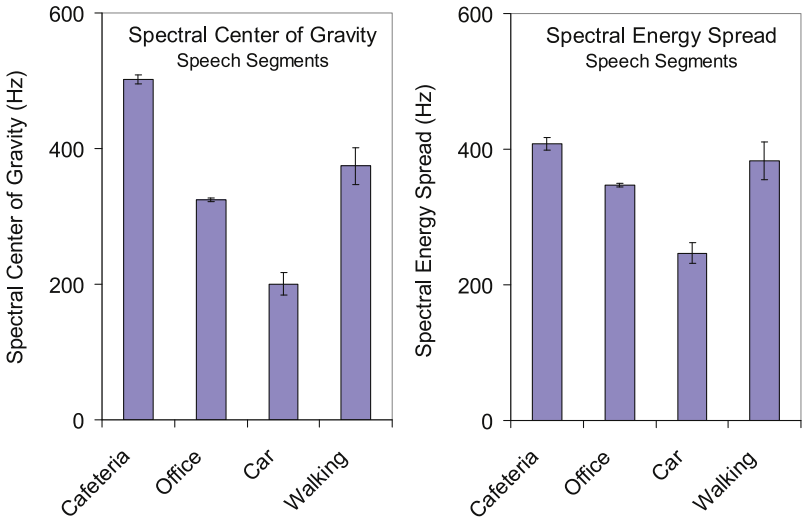

Fig. 20. Prof-Life-Log—spectral center of gravity and spectral energy spread in primary speaker in varying environments; error bars—95% confidence intervals.

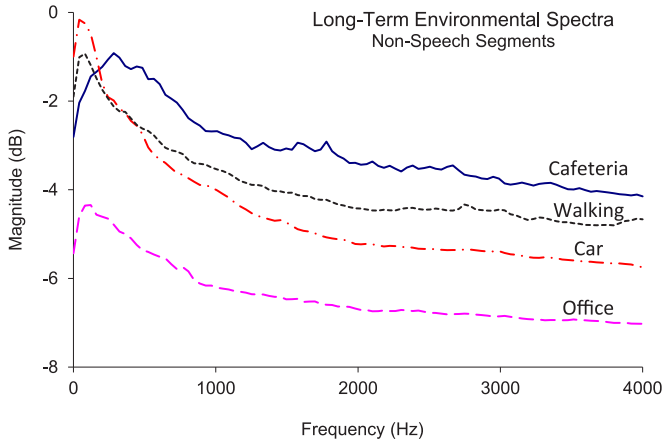

Fig. 21. Long-term spectra of selected environmental scenarios in Prof-Life-Log.

Figs. 17–19 analyze interactions between the vocal intensity in the primary speaker and the level and type of environmental noise (Lombard function), and also relationship between the primary and secondary speakers' average fundamental frequencies, shifts of vowels in the F_1–F_2 formant space, and speech rates as functions of the environment.

Fig. 17 details how the speakers cope with different environments in terms of vocal intensity (left—primary speaker) and fundamental frequency (right—primary speaker and secondary speakers). As could be expected, the naturally noisier environments *Cafeteria, Car* are seen to induce greater vocal intensities in the primary speaker compared to *Office* and *Walking*. On the other hand, fundamental frequency does not completely follow the same trend, with *Walking* fundamental frequency values exceeding those of the *Car*. Interestingly, the mean fundamental frequencies in secondary speakers follow an identical trend as those in the primary speaker when switching between environments. This suggests that (i) the environment has a rather consistent impact on the communication parties and/or (ii) there is a prominent inter-speaker convergence.

Fig. 18 studies shifts of vowel formants in the primary speaker due to the varying environments; in particular, prominent F_1 changes in /uw/ and F_2 changes in /eh/ are observed.

A similar trend to the one in F_0 can be seen for the speech rate in Fig. 19 where the communication parties again respond in the same fashion to different environments. The speech rate is estimated using the algorithm introduced in Heinrich and Schiel (2011a) and represents the number of syllable-like islands detected from the energy contour per second. Fig. 20 further details the mean spectral center of gravity (related to perceptual "brightness" of speech) and spectral energy spread measured for the primary speaker across the environments. The results suggest a strong correlation between the spectral characteristics of speech and the environment in which it was produced.

Fig. 21 presents long-term spectra of the four environments, with the office noise having the lowest energy and the car noise displaying a characteristic peak at low frequencies and a quick decay at higher frequencies. Given these prominent spectral profile differences, it seems quite natural that speakers adjust their speech production differently with each environment in an effort to sustain intelligible communication.

Fig. 22 compares distribution of pitch bigrams for the primary speaker in *Office* and *Cafeteria* environments. The pitch bigram analysis follows the procedure introduced in Bořil et al. (2013) where each voiced island in speech is fit with a regression line. Slopes of the extracted regression lines are quantized into three categories (pitch pattern primitives)—rising, flat, and falling. Subsequently, *N*-grams of pattern primitives can be counted and used to quantify differences in pitch contours. Fig. 22 presents absolute frequencies of primitive bigrams. Since a primitive unigram can take on one of three categories, there are nine possible bigrams. The shape of the particular bigram is given by its unigram coordinates read from left to right. For example, the central bigram represents a flat pattern (i.e., a combination of two flat unigrams). The bigram in the bottom corner represents a rising–falling pitch pattern. Fig. 22 shows that in the Office environment the flat pattern is dominating in the primary speaker while in Cafeteria, the bigrams are distributed more uniformly, suggesting greater pitch modulations.

For a comparison, Fig. 23 shows pitch bigram patterns found in the vocalizations of 18–31 month old US-born toddlers (five males and three females) produced during their natural daily routines (Bořil et al., 2013). The recordings were collected using the same type of a portable LENA recorder as in the Prof-Life-Log dataset. Unlike in the adult speech where the flat pitch bigram is most prominent, the toddler vocalizations are characterized by dominating double-falling and double-rising bigrams, with the frequencies of the flat bigram and the other remaining bigrams being nearly uniformly distributed. This corresponds well with the intuition that toddlers and children in general tend to modulate their voices at a notably higher rate than adults.

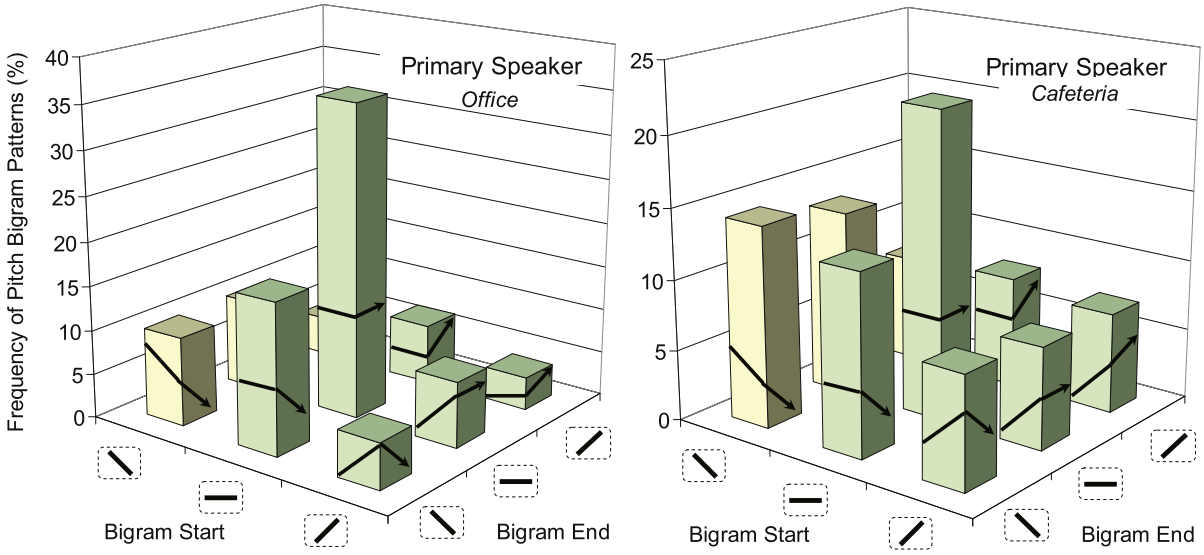

Fig. 22. Prof-Life-Log—distribution of pitch bigram patterns for primary speaker in Office and Cafeteria environments.

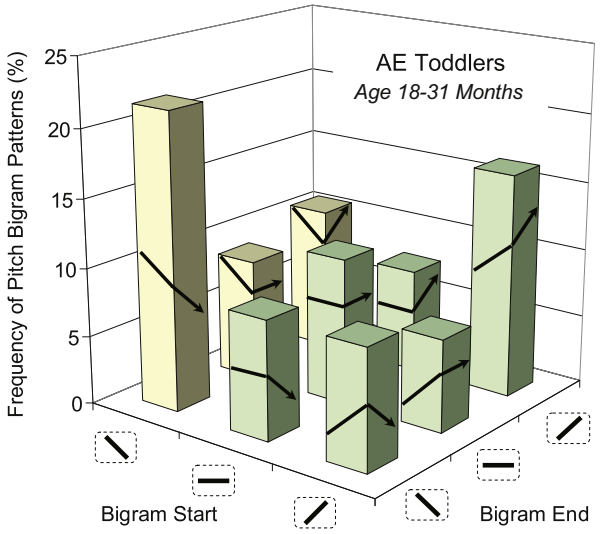

Fig. 23. Distribution of pitch bigram patterns in eight US toddlers aged 18–31 months.

6. Conclusions

This study has considered a range of prominent factors that impact speech and language technology. It should be noted that the emphasis here has been more on factors that impact speaker variability, which by no means is exhaustive, and therefore further investigations are clearly possible into sub-areas within the environment, channel, microphone/recording, context issues, etc. It is clear that the increasing extensive sub-areas within environment, channel, microphone/recording, context issues, etc. which also deserve further consideration. It is clear that the increasing demands placed on current speech engines have driven many speech researchers to leverage found data as a means to address naturalistic audio streams for speech recognition, speaker recognition, language identification, and diarization technology in general. While a variety of factors influence the presence or lack of realism in speech corpora as well as their corresponding effects on speech systems, for present day solutions to be effective, more realistic data is needed. In particular, our focus in this study has been on the role of speaker variability, environmental noise, communication factors, and channel variability. Here we have presented research examples of widely used corpora that in some ways departed from what could be considered realistic scenarios—either through the data collection protocol or due to the misinterpretation of the purpose of the data sets by users, alongside with data sets that we believe have the potential to address those issues.

In terms of recommendations for researchers going forward who want to address robustness in speech and language systems, the following three “best practices” are suggested:

- *Recommendation #1—Know your Source Plan*: for the specific speech or language application, know the number of speakers, amount of speech per speaker, speaking style (stress, emotion, vocal effort, Lombard effect, etc.) which is needed or acceptable for your application. Know if the audio streams are single speaker or multiple speakers, and if there is any overlap when speech is present. Know the language and dialect of interest—and recognize that code-switching occurs in bilingual speakers in the context of conversational subjects that are bilingual. Building an acoustic model at some level is like baking a cake—you need to have the right mix of speech ingredients, but too much of one style or context can alter the resulting model and degrade overall system performance.
- *Recommendation #2—Prune What is Not Needed*: developing screening tools to perform data purification or balancing is a must when dealing with found or massive audio streams or corpora. Setting aside data that does not meet the requirements set forth in the Source Plan is important and new signal processing tools are clearly needed to accomplish this task. This includes automatic SNR, quality, noise, and non-linear distortion analysis and visualization to build an “extrinsic” profile of the available audio materials. An equivalent “intrinsic” speaker profiling response is also needed to allow designers to construct the needed training materials to achieve the best acoustic models for system development. Determining effective thresholds or criteria to prune available corpora is an emerging and critical need for the field.
- *Recommendation #3—The need for Data Profiling & Proof-of-Concept Testing*: The final recommendation from this investigation is that the availability of speech and language resources continue to grow exponentially, and while using *found* data is tempting because it is cost effective, researchers need to exercise greater caution in how they treat such data. As such, there is a clear need for developing a “Data Profiling” paradigm which can be applied to both training data as well as sample test data. The need here is to independently measure the data diversity of both train and sample test to ensure there is a similar balance. To accomplish this, a series of basic analysis steps

for the training data should always include assessing: (i) SNR/noise level, (ii) channel characteristics, (iii) reverberation traits/room impulse response, (iv) number of speakers in the audio, (v) potential overlap of speech segments, (vi) presence of music or non-speech events/content, (vii) language of the speaker/speech, and (viii) identity and style of the speaker. Finally, a "proof-of-concept" test is needed. For example, in speaker or language recognition, one can always perform the evaluation on the speech after a SAD (speech activity detection) step is employed. A good proof-of-concept test here would be to test all the data which was set aside as silence/noise. If the models are truly balanced, the SID or LID task should provide roughly chance performance (e.g., 75% EER or 25% accuracy for a 4-way classification task).

Greater care in preliminary assessment of the audio content for training as well as probe testing can significantly increase the effectiveness and reliability of the final speech algorithm solution. New corpora are being made available to the research community with examples such as (i) SITW—speakers in the wild (useful for speaker ID), (ii) Fearless Steps Corpus (Yu and Hansen, 2017; Kaushik et al., 2017)—naturalistic audio from the U.S. NASA Apollo program (19,000 h; useful for diarization, speaker identification, speech under stress, etc.), (iii) NIST SRE and LRE corpora collected by NIST for speaker and language recognition evaluations (NIST, 2016; 2017), etc. While this study has considered a number of sub-areas in the field of speech, language, and speaker processing for recognition, it was not possible to provide an exhaustive treatment for all topic areas in robust speech research. Clearly, the next generation of speech and language solutions will need to take greater advantage of naturalistic data as well as designed data collections which meet the needs of researchers and technologists for robust solutions to the problems of speaker, environment, and communications based mismatch or variability.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.specom.2018.05.004.

References

Adank, P., Smits, R., van Hout, R., 2004. A comparison of vowel normalization procedures for language variation research. J. Acoust. Soc. Am. 116 (5), 3099–3107.

Akbacak, M., Hansen, J.H.L., 2007. Environmental sniffing: noise knowledge estimation for robust speech systems. IEEE Trans. Audio Speech Lang. Process 15 (2), 465–477.

Akbacak, M., Vergyri, D., Stolcke, A., Scheffer, N., Mandal, A., 2011. Effective Arabic dialect classification using diverse phonotactic models. Proceedings of the ISCA INTERSPEECH. Florence, Italy.

Angkititrakul, P., Petracca, M., Sathyanarayana, A., Hansen, J.H.L., 2007. UTDrive: driver behavior and speech interactive systems for in-vehicle environments. Proceedings of the IEEE Intelligent Vehicles Symposium. pp. 566–569.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems. Digit Signal Process. 10 (1–3), 42–54.

Baghel, S., Prasanna, S.R.M., Guha, P., 2017. Classification of multi speaker shouted speech and single speaker normal speech. Proceedings of the IEEE Region 10 Conference TENCON. pp. 2388–2392.

Banks, B., Gowen, E., Munro, K.J., Adank, P., 2015. Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. Front. Hum. Neurosci. 9, 422.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines. Proceedings of the IEEE ASRU. Scottsdale, AZ.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2017. The third CHiME speech separation and recognition challenge: analysis and outcomes. Comput. Speech Lang. 46, 605–626.

Biadsy, F., Hirschberg, J., Ellis, D.P.W., 2011. Dialect and accent recognition using phonetic-segmentation supervectors. Proceedings of the of ISCA INTERSPEECH. Florence, Italy, pp. 745–748.

Biadsy, F., Hirschberg, J., Habash, N., 2009. Spoken Arabic dialect identification using phonotactic modeling. Proceedings of the EACL Workshop on Computational Approaches to Semitic Languages. Athens, Greece, pp. 53–61.

Bond, Z., Moore, T., 1990. A note on loud and Lombard speech. Proceedings of the ICSLP. Kobe, Japan, pp. 969–972.

Bond, Z.S., Moore, T.J., Gable, B., 1989. Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask. J. Acoust. Soc. Am. 85 (2), 907–912.

Bordia, P., 1997. Face-to-face versus computer-mediated communication: a synthesis of the experimental literature. J. Bus. Commun. 1973 34 (1), 99–118.

Bořil, H., 2005. Automatic reconstruction of utterance boundaries time marks in speech database re-grabbed from DAT recorder. Proceedings of the of International Workshop on Digital Technologies. Zilina, Slovakia, pp. 13–16.

Bořil, H., Fousek, P., 2006. Influence of different speech representations and HMM training strategies on ASR performance. Acta Polytechnica, J. Adv. Eng. 46 (6), 32–35.

Bořil, H., Hansen, J.H.L., 2011. UT-Scope: towards LVCSR under Lombard effect induced by varying types and levels of noisy background. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, pp. 4472–4475.

Bořil, H., Hansen, J.H.L., Xu, D., Gilkerson, J., Richards, J., 2011. A longitudinal study of infant speech production parameters: a case study. Proceedings of the LENA Users Conference. Denver, Colorado.

Bořil, H., Pollák, P., 2005. Comparison of three Czech speech databases from the standpoint of Lombard effect appearance. Proceedings of the ASIDE, COST278 Final Workshop and ISCA Tutorial and Research Workshop. Aalborg, Denmark.

Bořil, H., Sadjadi, O., Kleinschmidt, T., Hansen, J.H.L., 2010. Analysis and detection of cognitive load and frustration in drivers' speech. Proceedings of the ISCA INTERSPEECH. Makuhari, Chiba, Japan, pp. 502–505.

Bořil, H., Sangwan, A., Hansen, J.H.L., 2012. Arabic dialect identification—'Is the secret in the silence?' and other observations. Proceedings of the ISCA INTERSPEECH. Portland, Oregon, pp. 30–33.

Bořil, H., Zhang, Q., Angkititrakul, P., Hansen, J.H.L., Xu, D., Gilkerson, J., Richards, J.A., 2013. A preliminary study of child vocalization on a parallel corpus of US and Shanghainese toddlers. Proceedings of the ISCA INTERSPEECH. Lyon, France, pp. 2405–2409.

Bou-Ghazale, S.E., Hansen, J.H.L., 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. Speech Audio Process. 8 (4), 429–442.

Bořil, H., 2008. Robust Speech Recogniton: Analysis and Equalization of Lombard Effect in Czech Corpora. Czech Technical University in Prague Ph.D. thesis. Czech Republic, http://www.utdallas.edu/~hxb076000

Bořil, H., Hansen, J.H.L., 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Trans. Audio Speech Lang. Process. 18 (6), 1379–1393.

Claes, T., Dologlou, I., ten Bosch, L., van Compernolle, D., 1998. A novel feature transformation for vocal tract length normalization in automatic speech recognition. IEEE Trans. Speech Audio Process. 6 (6), 549–557.

Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am. 128 (4), 2059–2069.

Cummings, K., Clements, M., 1990. Analysis of glottal waveforms across stress styles. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1. Albuquerque, USA, pp. 369–372.

Dreher, J.J., O'Neill, J., 1957. Effects of ambient noise on speaker intelligibility for words and phrases. J. Acoust. Soc. Am. 29 (12), 1320–1323.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1. pp. 346–348.

Eklund, I., Traunmuller, H., 1997. Comparative study of male and female whispered and phonated versions of the long vowels of swedish. Phonetica 1–21.

ELRA, 2008. European language resources association: SPEECON databases. URL: http://catalog.elra.info.

Fan, X., Hansen, J.H.L., 2010. Acoustic analysis for speaker identification of whispered speech. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5046–5049.

Fan, X., Hansen, J.H.L., 2011. Speaker identification within whispered speech audio streams. IEEE Trans. Audio Speech Lang. Process. 19 (5), 1408–1421.

Fan, X., Hansen, J.H.L., 2013. Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. Speech Commun. 55 (1), 119–134.

Frederic Aman, S.R., Michel Vacher, Portet, F., 2013. Analysing the performance of automatic speech recognition for ageing voice: does it correlate with dependency level? Proceedings of the 4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT). Grenoble, France, pp. 9–15.

Galic, J., Jovicic, S.T., Grozdic, D., Markovic, B., 2014. Constrained lexicon speaker dependent recognition of whispered speech. Proceedings of the Symposium on Industrial Electronics INDEL. pp. 180–184.

Garnier, M., 2007. Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal [Communication in noisy environments: From adaptation to vocal straining]. University of Paris VI, France Ph.D. thesis.

Ghaffarzadegan, S., Bořil, H., Hansen, J.H.L., 2014. Model and feature based compensation for whispered speech recognition. Proceedings of the ISCA INTERSPEECH. Singapore, pp. 2420–2424.

Ghaffarzadegan, S., Bořil, H., Hansen, J.H.L., 2015. Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia, pp. 5024–5028.

Ghaffarzadegan, S., Bořil, H., Hansen, J.H.L., 2016. Generative modeling of pseudo-whisper for robust whispered speech recognition. IEEE Trans. Audio Speech Lang. Process. 24 (10), 1705–1720.

Ghaffarzadegan, S., Bořil, H., Hansen, J.H.L., 2017. Deep neural network training for

whispered speech recognition using small databases and generative model sampling. Int. J. Speech Technol. 20 (4), 1063–1075.
Gramming, P., Sundberg, S., Ternström, S., Perkins, W., 1987. Relationship between changes in voice pitch and loudness. STL-QPSR 28 (1), 39–55.
Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G., Godfrey, J., 2010. Human assisted speaker recognition in NIST SRE10. Proceedings of the Odyssey: The Speaker and Language Recognition Workshop. pp. 180–185.
Grozdic, D.T., Jovicic, S.T., Galic, J., Markovic, B., 2014. Application of inverse filtering in enhancement of whisper recognition. Proceedings of the Symposium on Neural Network Applications in Electrical Engineering (NEUREL). pp. 157–162.
Hanilci, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., Ertaş, F., 2013. Speaker identification from shouted speech: Analysis and compensation. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8027–8031.
Hansen, J.H.L., 1988. Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition. Georgia Institute of Technology Ph.D. thesis. Atlanta, GA
Hansen, J.H.L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Commun. 20 (1–2), 151–173.
Hansen, J.H.L., Bořil, H., 2016. Robustness in speech, speaker, and language recognition: “You’ve got to know your limitations”. Proceedings of the ISCA INTERSPEECH. San Francisco, CA, pp. 2766–2770.
Hansen, J.H.L., Bria, O., 1990. Lombard effect compensation for robust automatic speech recognition in noise. Proceedings of the ICSLP. Kobe, Japan, pp. 1125–1128.
Hansen, J.H.L., Nandwana, M.K., Shokouhi, N., 2017. Analysis of human scream and its impact on text-independent speaker verification. J. Acoust. Soc. Am. 141 (4), 2957–2967.
Hansen, J.H.L., Ruzanski, E., Bořil, H., Meyerhoff, J., 2012. TEO-Based speaker stress assessment using hybrid classification and tracking schemes. Int. J. Speech Technol. 1–17.
Hansen, J.H.L., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, E.J., Anderson, T., Vloeberghs, C.R., Trancoso, I., Verlinde, P., 2000. The impact of speech under ‘stress’ on military speech technology. NATO Project Report.
Hansen, J.H.L., 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 (2), 366–378.
Hansen, J.H.L., Williams, K., Bořil, H., 2015. Speaker height estimation from speech: fusing spectral regression and statistical acoustic models. J. Acoust. Soc. Am. 138 (2), 1052–1067.
Hasan, T., Bořil, H., Sangwan, A., Hansen, J.H.L., 2013a. Multi-modal highlight generation for sports videos using an information-theoretic excitability measure. EURASIP J. Adv. Signal Process. 2013 (173), 1–17.
Hasan, T., Sadjadi, O., Gang, L., Shokouhi, N., Bořil, H., Hansen, J.H.L., 2013b. CRSS systems for 2012 NIST Speaker Recognition Evaluation. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, pp. 6783–6787.
Hatano, H., Kitamura, T., Takemoto, H., Mokhtari, P., Honda, K., Masaki, S., 2012. Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five japanese vowels uttered by fifteen male speakers. Proceedings of the ISCA INTERSPEECH. pp. 402–405.
Heeren, W.F.L., Lorenzi, C., 2014. Perception of prosody in normal and whispered french. J. Acoust. Soc. Am. 135 (4), 2026–2040.
Heinrich, C., Schiel, F., 2011. Estimating speaking rate by means of rhythmicity parameters. Proceedings of the ISCA INTERSPEECH. pp. 1873–1876.
Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., Okuno, H.G., 2015. Automatic speech recognition for mixed dialect utterances by mixing dialect language models. IEEE Trans. Audio Speech Lang. Process. 23 (2), 373–382.
Hirsch, H.G., Finster, H., 2005. The simulation of realistic acoustic input scenarios for speech recognition systems. Proceedings of the ISCA INTERSPEECH. pp. 2697–3000. Lisboa, Portugal.
Ito, T., Takeda, K., Itakura, F., 2001. Acoustic analysis and recognition of whispered speech. Proceedings of the IEEE ASRU. pp. 429–432.
Ito, T., Takeda, K., Itakura, F., 2005. Analysis and recognition of whispered speech. Speech Commun. 45 (2), 139–152.
Jin, Q., Schultz, T., Waibel, A., 2007. Far-field speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (7), 2023–2032.
Jou, S.-C., Schultz, T., Waibel, A., 2005. Whispery speech recognition using adapted articulatory features. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1. pp. 1009–1012.
Junqua, J., Anglade, Y., 1990. Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2. Albuquerque, USA, pp. 841–844.
Junqua, J.-C., 1993. The lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93 (1), 510–524.
Junqua, J.-C., 2002. Sources of Variability and Distortion in the Communication Process. Robust Speech Recognition in Embedded Systems and PC Applications. The International Series in Engineering and Computer Science. vol. 563. Springer US, Boston, MA, pp. 1–36.
Junqua, J.-C., Fincke, S., Field, K., 1998. Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition. Proceedings of the ICSLP. Sydney, Australia.
Kaushik, L., Sangwan, A., Hansen, J.H.L., 2017. Multi-channel Apollo mission speech transcript calibration. Proceedings of the ISCA INTERSPEECH. Stockholm, Sweden, pp. 2799–2803.
Kelly, F., Hansen, J.H.L., 2016. Score-aging calibration for speaker verification. IEEE Trans. Audio Speech Lang. Process. 24 (12), 2414–2424.
Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (4), 1435–1447.
Kent, R., Read, C., 2002. The Acoustic Analysis of Speech. Singular/Thomson Learning.
Kumar, K., Raj, R.S.B., Stern, R.M., 2011. Gammatone sub-band magnitude-domain dereverberation for ASR. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5448–5451.
Lane, H., Tranel, B., 1971. The lombard sign and the role of hearing in speech. J. Speech and Hear. Res. 14, 677–709.
Lee, P.X., Wee, D., Toh, H.S.Y., Lim, B.P., Chen, N., Ma, B., 2014. A whispered Mandarin corpus for speech technology applications. Proceedings of the ISCA INTERSPEECH. Singapore, pp. 1598–1602.
Lee, T., Liu, Y., Huang, P.W., Chien, J.T., Lam, W.K., Yeung, Y.T., Law, T.K.T., Lee, K.Y.S., Kong, A.P.H., Law, S.P., 2016. Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6475–6479.
Leongomez, J.D., Mileva, V.R., Little, A.C., Roberts, S.C., 2017. Perceived differences in social status between speaker and listener affect the speaker’s vocal characteristics. PLoS ONE 12, 1–21.
Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. 22 (4), 745–777.
Lim, B.P., 2011. Computational Differences between Whispered and Non-whispered Speech. University of Illinois at Urbana-Champaign Ph.D. thesis.
Liu, G., Hansen, J.H.L., 2014. An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios. IEEE Trans. Audio Speech Lang. Process. 22 (12), 1978–1992.
Lombard, E., 1911. Le signe de l’elevation de la voix. Ann. Malad. Oreille, Larynx, Nez, Pharynx 37, 101–119.
Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble and stationary noise. J. Acoust. Soc. Am. 124 (5), 3261–3275.
Lu, Y., Cooke, M., 2009. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun. 51 (12), 1253–1262.
Mathur, A., Reddy, S.M., Hegde, R.M., 2012. Significance of parametric spectral ratio methods in detection and recognition of whispered speech. EURASIP J. Adv. Signal Process. 2012 (1), 1–20.
Matsuda, M., Kasuya, H., 1999. Acoustic nature of the whisper. Proceedings of the EUROSPEECH. pp. 133–136.
Mehrabani, M., Hansen, J.H.L., 2013. Singing speaker clustering based on subspace learning in the GMM mean supervector space. Speech Commun. 55 (5), 653–666.
Mirsamadi, S., Hansen, J.H.L., 2016. A generalized nonnegative tensor factorization approach for distant speech recognition with distributed microphones. IEEE Trans. Audio Speech Lang. Process. 24 (10), 1721–1731.
Morris, R.W., Clements, M.A., 2002. Reconstruction of speech from whispers. Med. Eng. Phys. 24 (7), 515–520.
Najafian, M., Safavi, S., Hanani, A., Russell, M., 2014. Acoustic model selection using limited data for accent robust speech recognition. Proceedings of the European Signal Processing Conference (EUSIPCO). pp. 1786–1790.
Nandwana, M.K., Bořil, H., Hansen, J.H.L., 2015. A new front-end for classification of non-speech sounds: a study on human whistle. Proceedings of the ISCA INTERSPEECH. pp. 1982–1986. Dresden, Germany
Narayan, C.R., McDermott, L.C., 2016. Speech rate and pitch characteristics of infant-directed speech: longitudinal and cross-linguistic observations. J. Acoust. Soc. Am. 139 (3), 1272–1281.
NIST, 2016. Speaker recognition evaluation (SRE). URL: https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016.
NIST, 2017. Nist language recognition evaluation (LRE). URL: https://www.nist.gov/itl/iad/mig/nist-2017-language-recognition-evaluation.
Pardo, J., 2013. Measuring phonetic convergence in speech production. Front. Psychol. 4 (559), 1–5.
Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., 2004. Performance analysis of the Aurora large vocabulary baseline system. Proceedings of the European Signal Processing Conference (EUSIPCO). pp. 553–556.
Pearce, D., Hirsch, H.-G., Gmbh, E.E.D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proceedings of the ISCA ITRW ASR. pp. 29–32.
Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M.S., Braga, D., 2012. Impact of age in ASR for the elderly: preliminary experiments in european portuguese. In: Torre Toledano, D., Ortega Giménez, A., Teixeira, A., González Rodríguez, J., Hernández Gómez, L., San Segundo Hernández, R., Ramos Castro, D. (Eds.), Advances in Speech and Language Technologies for Iberian Languages. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 139–147.
Pick, H.L., Siegel, G.M., Fox, P.W., Garber, S.R., Kearney, J.K., 1989. Inhibiting the Lombard effect. J. Acoust. Soc. Am. 85 (2), 894–900.
Pisoni, D., Bernacki, R., Nusbaum, H., Yuchtman, M., 1985. Some acoustic-phonetic correlates of speech produced in noise. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 10. Tampa, Florida, pp. 1581–1584.
Ravanelli, M., Brakel, P., Omologo, M., Bengio, Y., 2017. A network of deep neural networks for distant speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4880–4884.
Rose, R.C., Hofstetter, E.M., Reynolds, D.A., 1994. Integrated models of signal and background with application to speaker identification in noise. IEEE Trans. Speech Audio Process. 2 (2), 245–257.

Ruff, S., Bocklet, T., Noth, E., Muller, J., Hoster, E., Schuster, M., 2017. Speech production quality of cochlear implant users with respect to duration and onset of hearing loss. ORL J. Otorhinolaryngol Relat Spec. 79 (5), 282–294.
Sadjadi, O.S., Bořil, H., Hansen, J.H.L., 2012. A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan, pp. 4701–4704.
Schulman, R., 1985. Dynamic and perceptual constraints of loud speech. J. Acoust. Soc. Am. 78 (S1), S37.
Sharifzadeh, H.R., McLoughlin, I.V., Russell, M.J., 2012. A comprehensive vowel space for whispered speech. J. Voice 26 (2), e49–e56.
Shokouhi, N., Hansen, J.H.L., 2017. Teager–kaiser energy operators for overlapped speech detection. IEEE Trans. Audio Speech Lang. Process. 25 (5), 1035–1047.
Sjolander, K., Beskow, J., 2000. WaveSurfer—an open source speech tool. Proceedings of the ICSLP. 4. Beijing, China, pp. 464–467.
Stone, M.A., Moore, B.C.J., 2002. Tolerable hearing aid delays ii. estimation of limits imposed during speech production. Ear. Hear. 23 (4), 325–338.
Sulyman, A., Bořil, H., Sangwan, A., Hansen, J.H.L., Ibiyemi, T.S., 2014. Engineering analysis and recognition of nigerian english: an insight into a low resource languages. Trans. Mach. Learn. Artif. Intell. 2 (3), 115–126.
Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: acoustic and perceptual analyses. J. Acoust. Soc. Am. 84 (3), 917–928.
Takizawa, Y., Hamada, M., 1990. Lombard speech recognition by formant-frequency-shifted LPC cepstrum. Proceedings of the ICSLP. Kobe, Japan, pp. 293–296.
Talkin, D., 1987. Speech formant trajectory estimation using dynamic programming with modulated transition costs. J. Acoust. Soc. Am. 82 (S1), S55.
Talkin, D., 1995. A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, Netherlands, pp. 495–518.
Tao, F., Busso, C., 2014. Lipreading approach for isolated digits recognition under whisper and neutral speech. Proceedings of the ISCA INTERSPEECH. Singapore, pp. 1154–1158.
Umesh, S., 2011. Studies on inter-speaker variability in speech and its application in automatic speech recognition. Sadhana 36 (5), 853–883.
Volín, J., Tykalová, T., Bořil, T., 2017. Stability of prosodic characteristics across age and gender groups. Proceedings of the ISCA INTERSPEECH. pp. 3902–3906.
Wagner, S., 2012. Age grading in sociolinguistic theory. Ling. Lang. Compass 6 (6), 371–382.
Webster, J.C., Klumpp, R.G., 1962. Effects of ambient noise and nearby talkers on a face-to-face communication task. J. Acoust. Soc. Am. 34 (7), 936–941.
Womack, B.D., Hansen, J.H.L., 1999. N-Channel hidden Markov models for combined stressed speech classification and recognition. IEEE Trans. Speech Audio Process. 7 (6), 668–677.
Xu, D., Gilkerson, J., Richards, J.A., 2012. Objective child vocal development measurement with naturalistic daylong audio recording. Proceedings of the ISCA INTERSPEECH. pp. 1123–1126.
Yang, C.-Y., Brown, G., Lu, L., Yamagishi, J., King, S., 2012. Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation. Proceedings of the Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 220–223.
Yu, C., Hansen, J.H.L., 2017. A study of voice production characteristics of astronuat speech during apollo 11 for speaker modeling in space. J. Acoust. Soc. Am. 141 (3), 1605–1614.
Zhang, C., Hansen, J.H.L., 2007. Analysis and classification of speech mode: whispered through shouted. Proceedings of the ISCA INTERSPEECH. pp. 2289–2292.
Zhang, C., Hansen, J.H.L., 2011. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. IEEE Trans. Audio Speech Lang. Process. 19 (4), 883–894.
Zhang, C., Yu, T., Hansen, J.H.L., 2010. Microphone array processing for distance speech capture: a probe study on whisper speech detection. Proceedings of the Asilomar Conference on Signals, Systems and Computers (ASILOMAR). pp. 1707–1710.
Zhang, Z., Weninger, F., Wöllmer, M., Han, J., Schuller, B., 2017. Towards intoxicated speech recognition. Proceedings of the International Joint Conference on Neural Networks (IJCNN). pp. 1555–1559.
Zhou, G., Hansen, J.H.L., Kaiser, J.F., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. 9 (3), 201–216.
Ziaei, A., Sangwan, A., Hansen, J.H.L., 2013. Prof-Life-Log: personal interaction analysis for naturalistic audio streams. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7770–7774.
Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Commun. 9, 351–356.