INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robustness in Speech, Speaker, and Language Recognition: "You've Got to Know Your Limitations"

John H. L. Hansen[*1], *Hynek Bořil*[1,2]

[1]Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
[2]Electrical Engineering Department, University of Wisconsin–Platteville, U.S.A.
John.Hansen@utdallas.edu, borilh@uwplatt.edu

Abstract

In the field of speech, speaker and language recognition, significant gains have and are being made with new machine learning strategies along with the availability of new and emerging speech corpora. However, many of the core scientific principles required for effective speech processing research appear to be drifting to the sidelines with the assumptions that access to larger amounts of data can address a growing range of issues relating to new speech/speaker/language recognition scenarios. This study focuses on exploring 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 mismatch/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/increased cognitive load conditions, and (iii) dialect identification using found data. Finally, we study speaker–noise 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.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

It is safe to say that the primary challenge in almost any speech, speaker or language processing/classification task is the ability to formulate a solution that overcomes mismatch between training and test conditions. Speech feature extraction, model training/development, and classification strategies have progressed significantly over the past fifty years, yet the overriding challenge continues to be the ability of speech/language algorithms to be *robust* as either speaker, technology/voice-capture, or environment based mismatch is introduced.

Why should speech researchers be concerned today? The primary reason is the overwhelming availability of *found* data in the field. The exponentially growing amount of speech data freely available causes a greater temptation to simply use whatever is available to address a specific research task. 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 they were originally collected for, they may in fact be constructing an irrelevant solution (e.g., [1]).

Figure 1 highlights three broad sources of mismatch: (i) speaker based (within or across speakers), (ii) conversation based, and (iii) technology/environment or noise based. Speaker-based variability (see Fig. 2) reflects a range of changes in how a speaker produces speech, and will impact system performance for either speech/speaker/language recognition. These can be thought of as intrinsic or within speaker variability, and include the following. *Situational Task Stress* – the subject is performing some task while speaking, such as operating a vehicle; hands-free voice input which can include cognitive [2, 3] as well as physical task stress [4]. *Vocal Effort/Style* – the subject alters their speech production from normal phonation, resulting in whisper [5–8] through shouted speech, or Lombard effect which occurs when the speaker produces speech in the presence of noise [9, 10]; or if they are singing vs. speaking [11]. *Emotion* – the subject is communicating their emotional state while speaking (e.g., anger, sadness, happiness, etc.) [12]. *Physiology* – effects of illness, intoxication, medication, and aging.

Conversation-based variability reflects voice interaction with either another person or technology, differences with respect to the language or dialect spoken [13], whether speech is read/prompted or spontaneous, or is a 2-way conversation or a public speech/group discussion.

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 [1, 14, 15]. *Environmental* – background noise [16] (stationary, impulsive, time-varying, etc.), room acoustics [17], reverberation [18, 19], distant microphone.

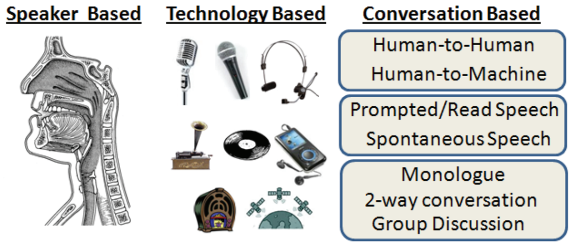

Figure 1: Mismatch in speech/language processing: (i) speaker, (ii) technology/environment/noise, (iii) conversation-based.

*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.

http://dx.doi.org/10.21437/Interspeech.2016-1395

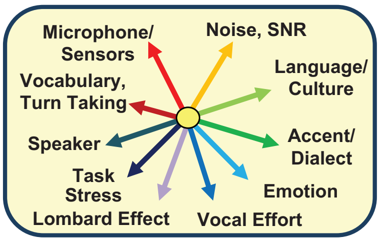

Figure 2: Sources of speaker-based variability.

Data quality – duration, sampling rate, recording quality, audio codec/compression [20].

Given the range of speaker, environment, acoustic, and technology based mismatch, what impact do these issues introduce to speech/speaker/language recognition systems, and what steps can researchers do to minimize these issues? The following sections explore several specific examples of mismatch due to noise, Lombard effect, communication scenario, emotions, and channel, and suggest when caution should be exercised.

2. Communication in Noise

In recent years, researchers have been putting a great deal of effort into the development of speech processing algorithms that would maintain good performance in real world conditions. Besides speaker/channel variability and room reverberation [19, 21], environmental noise represents one of the most disruptive and hard to deal with factors [22]. Successful modeling and suppression of noise effects in speech engines requires availability of noisy speech data. The most common approach to acquiring noisy data sets is to add noise to clean speech samples. This method is very flexible and economic as the same clean speech samples can be reused and mixed with different types of noise at desired signal-to-noise-ratios, without requiring a repeated participation of human subjects. This approach to noisy speech modeling has been taken, among others, by the creators of the popular Aurora datasets. In Aurora 2, noise was artificially added to clean recordings of TIDigits [23]; Aurora 4 followed the same concept with Wall Street Journal recordings [24], and Aurora 5 returned to TIDigits while expanding on simulated distortion factors that would include hands-free microphone, transmission through a GSM channel, and room reverberation [25]. Similar trend of creating challenging noisy datasets has been followed in other application domains (e.g., in the NIST Speaker Recognition campaigns [26]). The contribution of the Aurora suite is undisputed and quite remarkable – it has provided a unified development and evaluation framework for automatic speech recognition (ASR) and significantly accelerated the advancement of robust algorithms. This being said, as will be discussed below, the approach taken by Aurora and similar simulated noisy speech databases disregards the effects of noise on speech production.

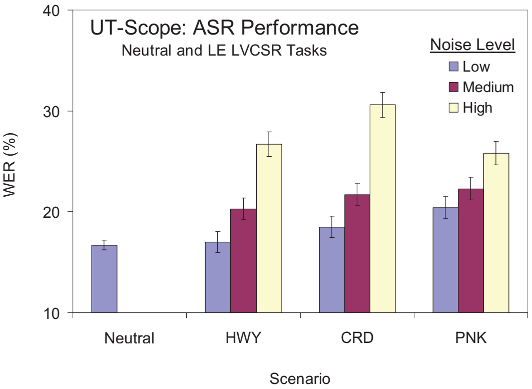

Figure 3: Talking in noise: ASR performance in clean neutral and clean Lombard speech UT-Scope tasks; TIMIT language model; 95% confidence intervals.

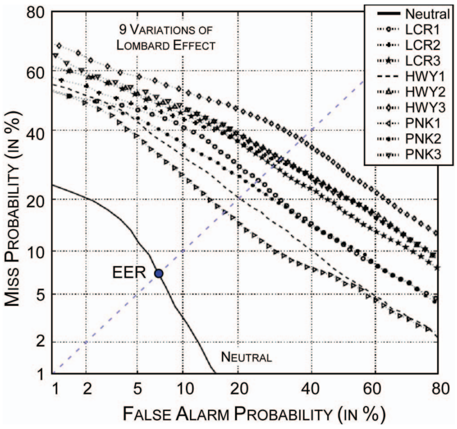

Figure 4: Talking in noise: SID performance in clean neutral and clean Lombard speech UT-Scope tasks; 12 sec. samples.

2.1. Adding Noise versus Talking in Noise

While mixing clean speech recordings with noise samples may provide a reasonable approximation to actual speech contamination by additive environmental noise, it does not reproduce the effects of noise on speakers. In reality, speakers continuously adjust their speech in response to the environmental noise to maintain intelligible communication (Lombard effect [4,27]). Lombard effect affects a number of speech production parameters [28–30]. Even if the additive noise is successfully removed or simply excluded from the Lombard speech recording, the speech variability due to Lombard effect may cause a severe mismatch with the neutral speech-trained acoustic models of a speech system and result in poor performance [31]. Figure 3

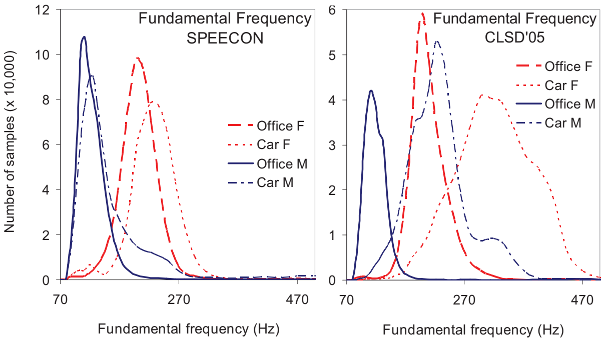

Figure 5: Talking in noise: fundamental frequency in scenarios without (SPEECON) and with (CLSD'05) communicaton factor; *F*/*M* – female/male subjects; 10 ms analysis window step.

demonstrates performance of a neutral speech-trained hidden Markov model (HMM) ASR system when tested on TIMIT-like [32] utterances produced by speakers that were exposed to three levels of a highway (HWY), large crowd (CRD/LCR), and pink noise (PNK) played back through headhpones (70, 80, and 90 dB SPL for HWY and CRD; 65, 75, 85 dB SPL for PNK). A close-talk microphone channel providing high SNR recordings was used in the ASR experiment on 31 US-born subjects' (25 females, 6 males) drawn from the UT-Scope Lombard Effect set [9] (see [33] for more details on the ASR experiment). It can be seen that the word error rate (WER) grows rapidly from the baseline no-noise *Neutral* condition once the speakers are exposed to increasing noise levels – while the recorded speech signal retains a high SNR.

Figure 4 presents DET curves for a speaker verification task (SID) on the UT-Scope database. Here, recordings from 30 subjects (19 males and 11 females) were used. Similar to the ASR task, the speech production changes induced by Lombard effect considerably deteriorate the system performance due to the increased mismatch with the reference speaker models (see [9] for more details). Clearly, none of the noise-induced speech changes discussed here could be observed in databases of neutral speech artificially mixed with noise (e.g., the Aurora sets [23–25]). In a consequence, it is unclear what the robustness of algorithms developed on such databases would be when exposed to speech produced in actual noise.

2.2. Talking in Noise: Reading versus Communicating

Some studies incorporate speech acquisition in realistic or simulated noisy conditions to address the issues discussed in the previous section. Yet, many of these efforts ask subjects to read aloud prompts in noise without providing them with any feedback whether their speech is intelligible to others [4,27,34,35]. In this way, there is no communication factor and the subjects simply read to themselves – their response to the noise being unpredictable. Communication factor plays a crucial role in naturalistic speech collection. The presence or lack of communication will result in significantly different speech characteristics [36]; an example is shown in Figures 5 and 6. The figures compare the fundamental frequency of speech and vowel locations in the F_1–F_2 formant plane for utterances from Czech SPEECON [37] and the Czech Lombard Speech Database (CLSD'05) [10]. In both cases, the subjects produced utterances in an office environment and when exposed to a car noise. In SPEECON, the subjects only read the prompts to themselves while in CLSD'05, they were asked to communicate the prompts to a listener exposed to the same noise, who would provide feedback on intelligibility. The rate of fundamental frequency and formant shifts is more prominent and at the same time also more consistent (more compact 1-σ ellipses) for the dataset involving the communication factor – CLSD'05 (see [10] for more details).

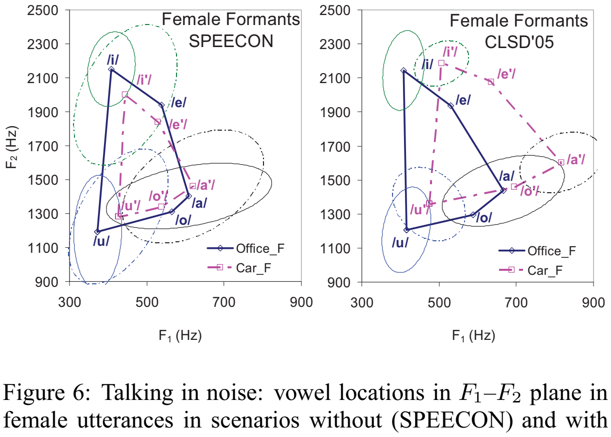

Figure 6: Talking in noise: vowel locations in F_1–F_2 plane in female utterances in scenarios without (SPEECON) and with (CLSD'05) communicaton factor; *F*/*M* – female/male subjects; 1-σ ellipses estimated to cover 39.4 % of samples.

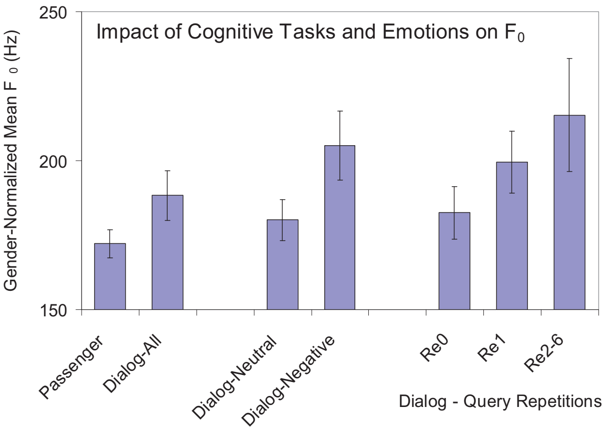

Figure 7: Talking in noise: human-human versus human-dialog system communication (UTDrive corpus); fundamental frequency; error bars represent 95% confidence intervals.

2.3. Talking in Noise: Communication Scenarios

This section presents an example from the UTDrive database [38] which utilizes recordings of 68 subjects driving in real traffic while performing various tasks. Figure 7 shows fundamental frequency estimated for the driver's speech while (i) casually talking to the passenger (*Passenger*), and (ii) calling a commercial dialog system with the task to find out a specific information (*Dialog*). Since the ASR portion of the dialog system produces frequent errors, the users had to often repeat their queries (*Re1–Re6*). The forced query repetitions might have negatively affected the drivers' mood – for this reason perceptual labels for *Neutral* and *Negative* states were extracted and studied alongside. The figure suggests that the communication mode (with a passenger or the dialog system), emotions (neutral/negative), and the number of query repetitions (1–6) all have impact on the speech production parameters. As discussed in more detail in [3], the production differences were prominent for a variety of speech parameters and provided a 94 % accuracy of distinguishing between the interactions with the dialog system versus the passenger when employed as features in a simple GMM-SVM classification scheme.

3. Channel Characteristics

Channel variability is usually viewed as a negative factor that increases mismatch between acoustic models and processed speech. Some speech corpora incorporate various channel characteristics to test robustness of speech engines (e.g., [25, 26, 35]). However, in some applications, channel characteristics may provide a valuable information about a particular environment or recording equipment used during the data acquisition. The left-hand side of Fig. 8 shows long-term channel characteristics estimated for LDC's conversational telephone speech

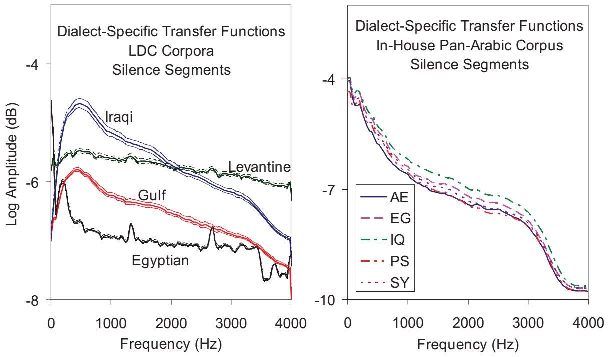

Figure 8: (Left) Dialect-specific channel characteristics in Arabic CTS corpora – dashed lines are ±5σ intervals; (right) channel characteristics in in-house Pan-Arabic corpus capturing dialects of United Arab Emirates (AE), Egypt (EGY), Iraq (IRQ), Palestine (PS), and Syria (SY).

(CTS) corpora capturing four Arabic dialects (Iraqi, Levantine, Gulf, Egyptian). As demonstrated in [1], one can perform a highly accurate dialect identification (DID) on these data sets using just silence segments since the channel characteristics are perfectly correlated with the respective dialects. This is an example of using a good data in a wrong context (see [39–41]). The CTS sets were originally collected for ASR purposes and were not intended for DID tasks. The right-hand side presents an in-house Pan-Arabic corpus that was collected with the DID task in mind using a fixed recording setup in all scenarios – note that the estimated channel characteristics for all dialect sets are nearly identical in this case. As shown in [1], a DID task on silence segments in this coprus yields a chance performance.

4. Prof-Life-Log: Naturalistic Speech Corpus

This section presents an example of a fully naturalistic corpus Prof-Life-Log [43] that contains audio recordings of entire work days in the life of the subject – a university professor.

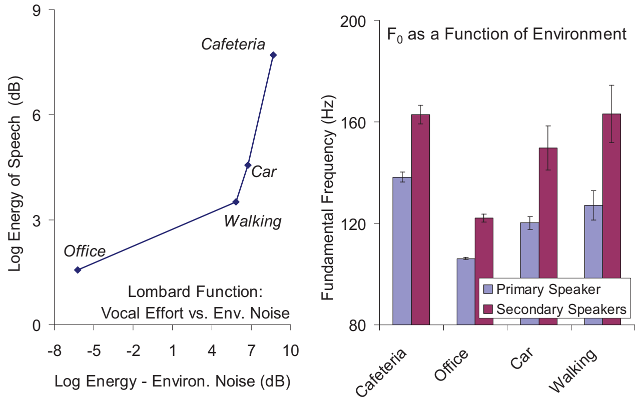

Figure 9: Prof-Life-Log: (Left) Lombard function – vocal intensity as a function of environmental scenario; primary speaker; (right) fundamental frequency in primary speaker and secondary speakers in varying environments; error bars – 95% confidence intervals.

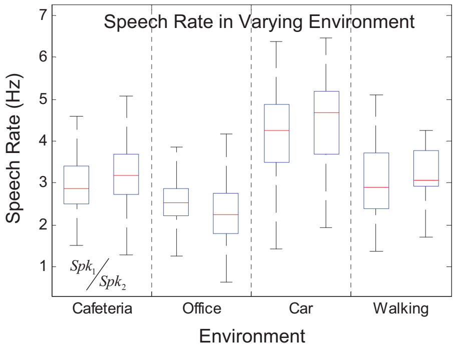

Figure 10: Prof-Life-Log: speech rate (extracted from short-time energy envelopes, following [42]) in primary and secondary speakers as a function of environment.

The recordings are acquired using the LENA (Language Environment Analysis) unit [44] and capture unscripted interactions of the professor (*primary speaker*) with colleagues, students, and acquaintances (*secondary speakers*) in various environments and provide a unique insight into natural speech communication. Figures 9 and 10 study the relationship between the primary spekar's vocal intensity and the level of environmental noise (Lombard function), and the estimated mean fundamental frequency and speech rate in the primary speaker versus averaged secondary speakers. The analyses were conducted on 12 hours of audio (one work day) capturing *Office*, *Cafeteria*, *Walking*, and *Car* environments. The trends in the figures capture combined human-environment and human-human interactions in natural communication. It can be seen that the primary speaker raises his vocal intensity and pitch when exposed to noisier environments (*Cafeteria*, *Car*, *Walking* versus *Office*). While the primary speaker has a lower than average nominal pitch, the averaged pitch of the secondary speakers follows the same trend across the environments, which may be a response to the environmental noise, combined with adjustments to the talking style of the primary speaker. Similar trend can be seen for the speech rate in Fig. 10.

5. Conclusions

This paper studied a variety of factors influencing the presence or lack of realism in speech corpora and the corresponding effects on speech systems. In particular, our focus was on the role of environmental noise, communication factor, and channel variability. We presented 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 in our belief have the potential to address those issues.

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. A series of basic analysis steps should always include assessing (i) SNR, (ii) number of speakers, (iii) potential overlap of speech segments, (iv) presence of music or non-speech events/content, and (v) identity and style of the speaker and their language. Greater care in preliminary assessment of the audio content can significantly increase effectiveness and reliability of the final speech algorithm solution.

6. References

[1] H. Bořil, A. Sangwan, and J. H. L. Hansen, "Arabic dialect identification - 'Is the secret in the silence?' and other observations," in *Proc. of INTERSPEECH 2012*, Portland, Oregon, pp. 30–33.

[2] J. Hansen, E. Ruzanski, H. Bořil, and J. Meyerhoff, "TEO-based speaker stress assessment using hybrid classification and tracking schemes," *Int. Journal of Speech Technology*, pp. 1–17, 2012.

[3] H. Bořil, O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers speech," in *Proc. of INTERSPEECH'10*, Sep 2010, pp. 502–505.

[4] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[5] X. Fan and J. H. L. Hansen, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Communication*, vol. 55, no. 1, pp. 119–134, 2013.

[6] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 883–894, May 2011.

[7] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Model and feature based compensation for whispered speech recognition," in *Proc. INTERSPEECH'14*, Singapore, Sept 2014, pp. 2420–2424.

[8] ——, "Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition," in *Proc. of IEEE ICASSP 2015*, Brisbane, Australia, April 2015, pp. 5024–5028.

[9] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. ASLP*, vol. 17, no. 2, pp. 366 –378, Feb. 2009.

[10] H. Bořil, "Robust speech recogniton: Analysis and equalization of Lombard effect in Czech corpora," Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, http://www.utdallas.edu/∼hxb076000, 2008.

[11] M. Mehrabani and J. H. L. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Comm.*, vol. 55, no. 5, pp. 653–666, 2013.

[12] J. H. L. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. Vloeberghs, I. Trancoso, and P. Verlinde, "The impact of speech under 'stress' on military speech technology," in *NATO Project Report*, 2000.

[13] S. Amuda, H. Bořil, A. Sangwan, J. H. L. Hansen, and T. S. Ibiyemi, "Engineering analysis and recognition of Nigerian English: An insight into a low resource languages," *Trans. Machine Learning and Artificial Intelligence*, vol. 2(3), pp. 115–126, 2014.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[15] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 13, pp. 42 – 54, 2000.

[16] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, Apr 1994.

[17] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2023–2032, Sept 2007.

[18] C. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri, G. Doddington, and J. Godfrey, "Human assisted speaker recognition in NIST SRE10," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, p. 180185.

[19] O. S. Sadjadi, H. Bořil, and J. H. L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort," in *IEEE ICASSP'12*, Kyoto, Japan, 2012, pp. 4701–4704.

[20] H. Bořil and P. Fousek, "Influence of different speech representations and HMM training strategies on ASR performance," *Acta Polytechnica, J. on Adv. Eng.*, vol. 46, no. 6, pp. 32–35, 2006.

[21] K. Kumar, R. S. B. Raj, and R. M. Stern, "Gammatone subband magnitude-domain dereverberation for ASR," in *Proc. IEEE ICASSP*, May 2011, pp. 5448–5451.

[22] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745 – 777, April 2014.

[23] D. Pearce, H.-G. Hirsch, and E. E. D. Gmbh, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000, pp. 29–32.

[24] N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," in *Proc. of EUSIPCO*, Sept 2004, pp. 553–556.

[25] H. G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. INTERSPEECH 2005*, Lisboa, Portugal, 2005, pp. 2697–3000.

[26] T. Hasan, O. Sadjadi, L. Gang, N. Shokouhi, H. Bořil, and J. H. L. Hansen, "CRSS systems for 2012 NIST Speaker Recognition Evaluation," in *IEEE ICASSP 2013*, Vancouver, Canada, May 2013, pp. 6783–6787.

[27] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, vol. 93(1), 1993.

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

[29] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble and stationary noise," *JASA*, vol. 124, no. 5, pp. 3261–3275, 2008.

[30] ——, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253 – 1262, 2009.

[31] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, August 2010.

[32] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm.*, vol. 9(4), 1990.

[33] H. Bořil and J. H. L. Hansen, "UT-Scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," in *IEEE ICASSP'11*, Prague, 2011, pp. 4472–4475.

[34] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, 2008.

[35] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE ASRU 2015*, Scottsdale, Dec 2015.

[36] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *J. Speech and Hear. Res.*, vol. 14, pp. 677–709, 1971.

[37] ELRA, "European language resources assoc.: Speecon databases," 2008. [Online]. Available: http://catalog.elra.info

[38] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *IEEE Intelligent Vehicles Symposium*, June 2007, pp. 566–569.

[39] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proc. of EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009, pp. 53–61.

[40] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proc. of INTERSPEECH'11*, Florence, Italy, 2011, pp. 745–748.

[41] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, , and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models," in *Proc. of INTERSPEECH'11*, Florence, Italy, 2011.

[42] C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters," in *Proc. of INTERSPEECH*, 2011.

[43] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Prof-Life-Log: Personal interaction analysis for naturalistic audio streams," in *Proc. of IEEE ICASSP'2013*, May 2013, pp. 7770–7774.

[44] D. Xu, J. Gilkerson, and J. A. Richards, "Objective child vocal development measurement with naturalistic daylong audio recording," in *Proc. of INTERSPEECH*, 2012.