ANALYSIS OF SPEECH AND LANGUAGE COMMUNICATION FOR COCHLEAR IMPLANT USERS IN NOISY LOMBARD CONDITIONS

Jaewook Lee, Hussnain Ali, Ali Ziaei, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, Richardson, TX, USA

{jaewook, hussnain.ali, ali.ziaei, john.hansen}@utdallas.edu

ABSTRACT

Acoustic/linguistic modification of speech production with respect to auditory feedback is an important research domain for robust human-to-human and human-to-machine communication. For instance, in the presence of environmental noise, a speaker experiences the well-known phenomenon termed as Lombard effect. Lombard effect has been well studied for normal hearing listeners as well as for automatic speech/speaker recognition systems. However, limited effort has been employed to study if the speech production of cochlear implant (CI) users is influenced by the auditory feedback. The purpose of this study is to analyze the speech production and natural language model of CI users with respect to environmental changes. A mobile personal audio recording from continuous single-session audio streams collected over an individual's daily life was used for our study. The findings from this study will provide fundamental knowledge on the characteristics of speech production under Lombard effect in CI users. These specific variations in speech production can be leveraged in new algorithm development and further applications in speech systems to benefit cochlear implant users.

Index Terms— Speech analysis, speaker variability, cochlear implant, Lombard effect, natural language model

1. INTRODUCTION

Lombard effect is well-known phenomenon a speaker experiences in the presence of noise [1-4]. This phenomenon is perceptually realized with an increase in vocal effort such as amplitude, fundamental frequency, or formant location, and helps to maintain speech intelligibility over challenging listening environments. It is well documented that Lombard effect not only affects the intelligibility in speech communication, but it also degrades speech technology such as automatic speech recognition (ASR) and speaker identification (SID) [5-8]. Although well studied for normal hearing listeners and automatic speech/speaker recognition, Lombard effect has received little attention in the field of cochlear prostheses.

A number of reports suggest that whether the speech production of cochlear implant users is under control of any form of auditory feedback. Some studies examined the short-time effect of auditory feedback on speech production, which is related to the Lombard effect [9, 10]. Svirsky and Tobey support that rapid change in formant frequencies of vowels produced by a single female user of a Nucleus multichannel implant was observed when turning speech processor either on or off [9]. It is also argued by Svirsky et. al [10] that a number of speech parameters, such as fundamental frequency and vowel duration, that are available within relatively short-time constraints (few seconds or less) demonstrated immediate response to the speech processor on-versus-off conditions. These findings, however, were established when auditory feedback is artificially distorted, thus do not necessarily provide information about speech production in real communication conditions, such as noisy environments.

The objective of this study is to analyze speech production of cochlear implant users with respect to environmental noise conditions. In addition, the study aims to investigate the effect of auditory feedback on speech production in naturalistic daily environments. We observe and study this effect using mobile personal audio recordings from continuous single-session audio streams collected over an individual's daily life. Prior advancements in this domain include the "Prof-Life-Log" longitudinal study at UT-Dallas [11, 12].

In this study, four CI users who were all post-lingual deafened adults participated by producing spontaneous speech in various naturalistic noisy environments located on college campus including: office, hallway, outdoors on campus, and college gameroom. A number of parameters that are sensitive to Lombard speech were measured from the speech. In this research, analysis of speech production was accomplished in three ways: (i) characteristics of background noise and listening environment, (ii) acoustic analysis of speech production, and (iii) word selection and natural language model analysis. For the first part, two approaches were selected for knowledge on real-world environments. These are long-term averaged spectra and signal-to-noise ratio. The second part of the analysis is regarding feature extraction of glottal voice source from the speech. This includes fundamental frequency and glottal spectral slop. Lastly, a number of measures for lexical/linguistic selection of each

This work was supported by Grant R01 DC010494-01A awarded from the National Institute on Deafness and other Communication Disorders (NIDCD) of the National Institutes of Health (NIH).

speaker were investigated in terms of different listening conditions. In this part, word rate, unique word rate, conversational turn rate and perplexity of language model were considered. These analyses mentioned above will help us to explore the dependency of variation of speech upon changing environmental conditions.

2. APPLICATIONS

Long-term personal audio recordings are to be the wide range of potential application for cochlear implant recipients. This type of recording contains an abundance of information regarding speaker, speech, environments, language, etc. In recent years, speech and language processing capabilities (ASR, SID, etc.) in conjunction with personal mobile computing devices (e.g., smartphone [13], google glasses [14]) have opened new doors for data mining. The most promising use of naturalistic audio streams for CI users is in analyzing language acquisition and development of infant and young children [15]. These analyses have been performed by measuring various metrics of interest, such as adult word count, adult-child turn taking, child vocalizations, etc. Another potential application of personal audio recording of CI recipients is in the use of screening and diagnosis procedures for speech-related disorders in early childhood [16]. For example, analysis of acoustic features from realistic audio data can provide the capability of differentiating children with and without disorders such as autism or language delay. Furthermore, the capability of audio environment detection in conjunction with appropriate environmental-optimized coding algorithm can also be of practical use in personal audio recordings of CI users. These user customized paradigms can help us tap into the full potential of existing CI devices, which are currently not optimized for either the individual patients or clinicians for different users/environments [17, 18].

3. CORPUS DEVELOPMENT

The corpus here is designed to capture audio from a subject's daily life and investigate the influence of auditory feedback on speech production in naturalistic settings. We use the LENA device [19] for collecting naturalistic audio from CI users in this study. The LENA is a lightweight compact digital audio recorder that is capable of capturing mono audio data continuously for up to 16 hours. Fig. 1(a) demonstrates how the device is positioned for collecting naturalistic audio data using LENA. For capturing speech signal, a cross pack which was made of meshed-material was used to hold the device inside a pocket for secure and consistent placement. The device was located at the center of the chest where it is stationary with respect to the subject's mouth. This makes it possible for the unit's microphone to detect the speech signal more robustly against environmental noise during data collection.

A total of four CI users (mean age: 65 yrs.) who were fitted with the Nucleus device from Cochlear Ltd. participated in this study. All CI users were post-lingually deaf (lost hearing after the age of 18) and used their cochlear implant devices for at least four years. A total of four normal hearing (NH) speakers (mean age: 37 yrs.) participated as a conversation partner of the CI participants. The CI speakers in this research acted as the primary speaker, while the NH listeners served as the secondary speaker/listener. Note that the objective of this study is to analyze the speech production of CI users.

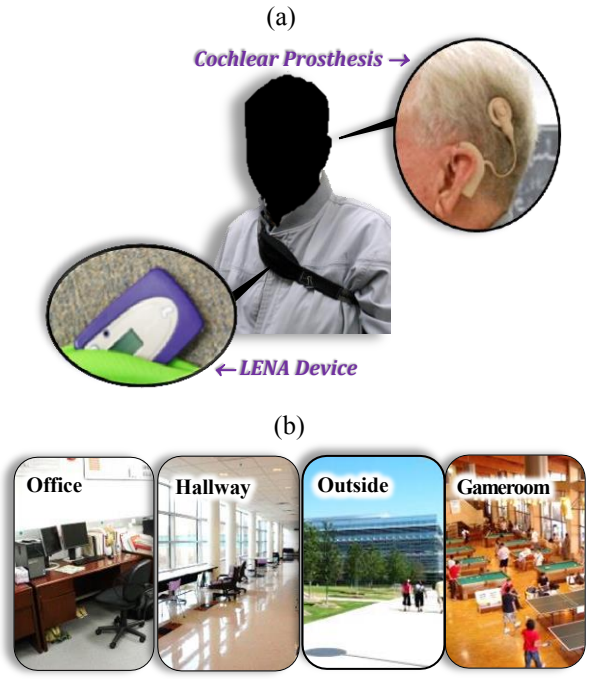

Figure 1: Naturalistic data collection for CI subject: (a) setup for data acquisition using the LENA unit, and (b) four locations on UTD campus for data collection.

Naturalistic audio was collected in 4 designated locations on a college campus: (i) office, (ii) hallway, (iii) outside on campus, and (iv) college gameroom, shown in Fig. 1(b). Noise conditions including type, mixture, and level varied greatly across the four conditions. In each location, prior to the subject's speech production, 3 minutes of background noise was recorded. These background noises were used to assess subjects' listening environments for subsequent analyses. Following the background recording, subjects were asked to perform free conversation between each other for 5 minutes in each location. A list of topics was provided to participants as a suggestion before the test, which included general topics, such as sports, news, weather, movies, etc. The subjects were noted that they are able to pause the audio recording anytime when they intended to or when privacy and confidential concerns arose during the recording.

Table 1: Text transcription analysis. All measurements except the conversation time were averaged from 4 CI individuals. The conversation time includes communication between the CI and NH participants.

	Conv. Time (sec.)	Speech Time (sec.)	Word Count	Unique Word Count	Conv. Turn Count
Office	310	132	472	196	88
Hallway	290	120	418	175	85
Outside	410	113	425	174	93
Gameroom	347	126	437	192	85

A set of acoustic labels were assigned to each audio track based on events in that space (i.e., sound events in the office space were different than outside in public areas). Every single utterances (sentence, phrase, word, and syllable) produced by CI speakers were identified manually based on listening audio. These labels were then used to compute acoustic parameters, removing the leading and trailing silent intervals. Additionally, acoustic environments (e.g., office, hallway, etc.) and speakers (e.g., SPK1, SPK2, etc.) labels were applied to the recordings. Finally orthographic transcripts for each utterance were created by human transcriber while listening to the audio. Here, only a single annotator was used for both acoustic and orthographic transcript labelling task for consistent evaluation. Table 1 summarizes general analysis of manual text transcripts obtained from the 4 CI users' communication.

4. ANALYSIS

Next in this section, we consider methods for analyzing production of conversational speech as a function of varying environment, and present results across the Lombard effect environments.

4.1. Noise/Environment Analysis

First, we analyzed background noise as well as subject's listening environment prior to analysis of speech production. To that end, we used two metrics: (i) long-term averaged spectrum, and (ii) signal-to-noise ratio (SNR). The long-term averaged spectrum was calculated using the pure background noise audio sampled during data collection. The SNR measures were simply obtained from subtracting the overall energy level of the background noise from that of speech production in dB scale. Two different approaches were employed to predict SNR levels with and without Lombard effect. While the SNR without Lombard effect was calculated using the office speech signals as the default clean reference for all locations, the SNR with Lombard effect was estimated using the speech signal with increased vocal intensity at each noisy environment. It has been shown in [7] that the type of noise as well as level causes the production of different "flavor" of Lombard effect. Note that in this study, we established the office environment as the quiet baseline, assuming that the speech produced in this location was neutral.

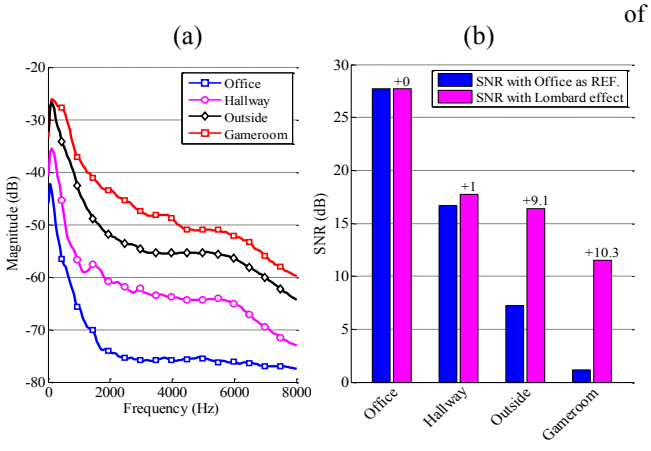

Figure 2: (a) Spectral characteristic of real-world maskers using the long-term averaged spectrum. (b) Evaluation of subject's listening environment using the signal-to-noise ratios (SNR) with and without Lombard effect respectively.

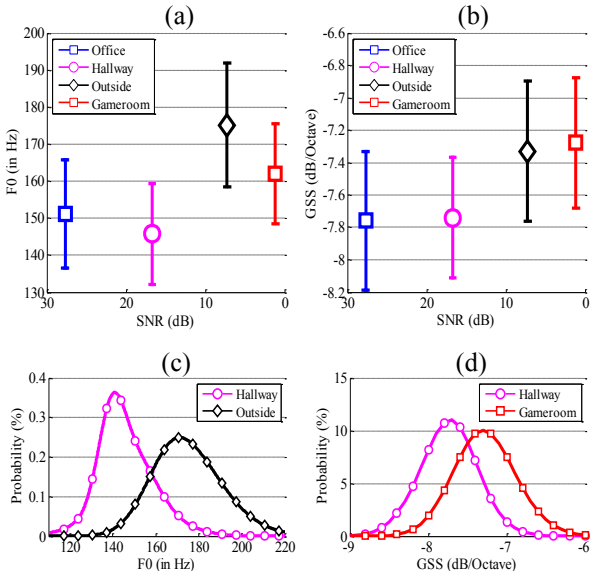

Figure 3: Acoustic analysis of speech production: Upper panel is the variation of the speech parameters, (a) fundamental frequency (F0) and (b) glottal spectral slope (GSS), as a function of varying SNR levels. Lower panel is the probability distributions of each parameter for the two most separated conditions.

Fig. 2(a) shows the long-term averaged spectrum for the four real-world maskers. It is shown from the figure that the office environment has the least spectral impact in overall spectral energy versus the other three environments. When compared to the office environment, significant increases in energy values were observed in most frequency bands, and this difference enhanced in moving from hallway to outside and gameroom conditions (i.e., approximately a 25dB increase in high frequency noise level from office to gameroom).

Fig. 2(b) shows the signal-to-noise ratios (SNRs) with respect to environments. In the figure, two bars on the left- and right-hand side for each condition correspond to the SNR with and without Lombard effect respectively. It clearly shows that the most favorable environment is office, and the most challenging scenario is gameroom in terms of speech quality. The SNR level without Lombard effect decreases by 27dB in office environment and down to 1dB when the environment switches to gameroom (left-side bar-graph plot). The decreased SNRs were, however, recovered due to the presence of Lombard effect by +1dB, +9.1dB and +10.3dB in hallway, outside, and gameroom respectively (right-side bar-graph plots). Therefore, the ability to include Lombard effect production within these environments provides from +1 and up to +10.3dB boost in SNR.

4.2. Speech Analysis

In order to examine the acoustic changes in speech production parameters, two matrices, fundamental frequency (F0) and glottal spectral slope (GSS) were considered [20-22]. F0 was computed

using Wavesurfer software [23]. GSS was computed from the glottal source spectrum of the speech signal over the frequency range from 500Hz to 2000Hz (2 octaves). A technique for estimating the glottal source spectrum has been adopted from Voicebox software [24]. The same analysis window of 20ms with 10ms frame increment was employed for both measurements at a 16 kHz sampling rate. Note that while some variability among individual subjects was observed, we consider average same conditions in this section.

In Fig. 3(a) and 3(b), the variation of F0 and GSS parameters are shown as a function of varying SNR levels. Note that SNR levels presented in each plot are measured assuming office speech signal as a clean default reference. In general, both parameters appear to be valid relayers for Lombard effect. For outside and gameroom conditions, the mean values for both speech parameters are shown to increase significantly compared to quiet baseline (office) as well as hallway condition.

Fig. 3(c) and 3(d) present two distributions for F0 and GSS parameters respectively to illustrate how much they are separated between each other. In these figures, the two most separated conditions for each parameter were selected, these are hallway and outside for F0 and hallway and gameroom for GSS. The results indicate that the distribution of each parameter varied significantly across conditions on campus. The relative increase between the two conditions were approximately +20% for F0 and +5.3% for GSS.

4.3. Language/Word Analysis

Lastly, four parameters that span the variation of word/linguistic structure in spontaneous speech were considered: (i) word rate (WR), (ii) unique word rate (UWR), (iii) conversational turn rate (TR), and (iv) word perplexity (PPL). The unique word rate refers to counts of unique individual words over time, while the word rate includes reoccurrences of the same words. These parameters were obtained based on analysis of manual text transcripts previously given in Table 1. The WR, UWR and TR were computed from the number of words, unique words, and conversational turns divided by CI's speech time. The PPL was estimated from the bi-gram natural language model trained using our corpus (CI's spontaneous speech) compared to the baseline language model trained using the Switchboard corpus (NH's spontaneous speech) [25, 26]. The creation and evaluation of the language models used in this test were supported via the capability of SRILM toolkit [27]. It is noted that since word selection and perplexity is highly speaker- and corpus-dependent, data for individual speakers were not combined in this section.

Fig. 4(a)-(d) show the distributions of WR, UWR, TR and PPL parameters respectively for 4 CI speakers across different environments. From the results, greater variation in each linguistic parameter across different environments is observed. In addition, large variability among the individual speakers was also seen in some cases. For example, significant larger shifts in SPK2 for UWR and SPK4 for PPL were observable when compared to the other three speakers in the location of outside. The dispersion of these parameters was found to be much higher relative to the other conditions. From these results, it is clearly seen that no consistent pattern of language/word shift over environments was found across

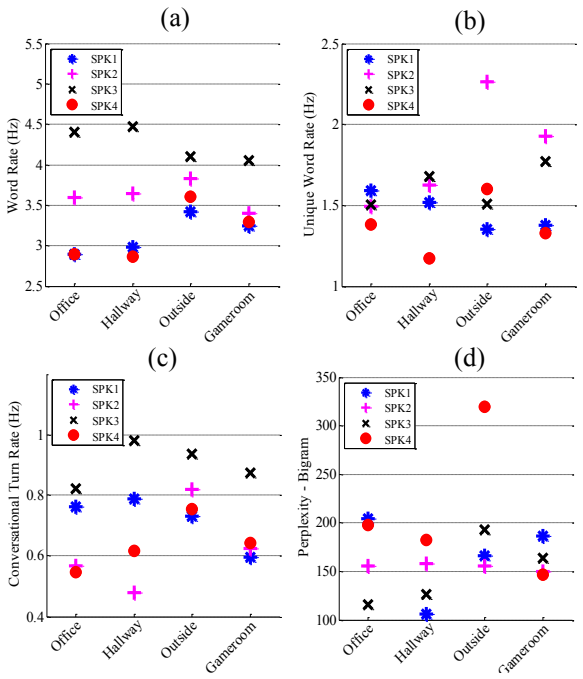

Figure 4: Word selection and natural language model analysis using (a) word rate, (b) unique word rate, (c) conversational turn rate and (d) perplexity for each individual CI users.

parameters even if there were large variation within different speakers/environments. This is due to word and language features are likely to be sensitive to speakers and corpus, rather than that of listening environments.

5. DISCUSSION AND CONCLUSIONS

In this study, we analyzed the speech production of CI users using mobile personal audio recordings collected over an individual's daily life. The results indicated that Lombard effect has been found in speech of cochlear implant users who are post-lingually deaf adults. Speakers increased their vocal efforts, such as fundamental frequency and glottal spectral slope significantly in challenging noisy environments to ensure intelligible communication in the presence of noise. In addition, it was observed that variation in word selection and language perplexity occurred within different speakers/environments, even if there is no consistent pattern of change across parameters.

The above result has the potential of playing an important role in further applications of speech and language technology especially for hearing impaired patients with cochlear implant. Historically, we know that different environments will have different noise types and levels. Traditional front-end processing for hearing aid and cochlear implants, for example, focus on noise suppression for minimizing the impact of noise, and are not optimized for different environments or users. Here, we have shown a fundamental shift in speech/language characteristics due to the Lombard effect in CI users. This change in speech production should be leveraged in new algorithm development and further applications of speech technology which are integrated for cochlear implant users.

6. REFERENCES

[1] E. Lombard, "Le signe de l'elevation de la voix [The sign of voice rasing]," *Annals Des Maladies De l'Oreille Et Du Larynx,* pp. 101-119, 1911.

[2] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech, Language and Hearing Research,* vol. 14, pp. 677-709, 1971.

[3] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," *Ph. D. Thesis, Georgia Institute of Technology, Atlanta, GA,* 1988.

[4] J. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Commun.,* vol. 20, pp. 13-22, 11, 1996.

[5] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. on Speech and Audio Proc.,* vol. 2, pp. 598-614, 10/01, 1994.

[6] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication,* vol. 20, pp. 151-173, 1996.

[7] 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. on Audio, Speech and Lang. Proc.,* vol. 17, pp. 366-378, 02/01, 2009.

[8] M. Saleem, G. Liu and J. H. L. Hansen, "Weighted training for speech under Lombard effect for speaker recognition," *ICASSP 2015,* 2015.

[9] M. A. Svirsky and E. A. Tobey, "Effect of different types of auditory stimulation on vowel formant frequencies in multichannel cochlear implant users," *Journal of the Acoustical Society of America,* vol. 89, pp. 2895-2904, 06/01, 1991.

[10] M. A. Svirsky, H. Lane, J. S. Perkell and J. Wozniak, "Effects of short-term auditory deprivation on speech production in adult cochlear implant users," *Journal of the Acoustical Society of America,* vol. 92, pp. 1284-1300, 09/01, 1992.

[11] A. Ziaei, A. Sangwan and J. H. L. Hansen, "Prof-Life-Log: Audio Environment Detection for Naturalistic Audio Streams," *Conference of the Int'l Speech Communication (INTERSPEECH'12),* pp. 2514-2517, 2012.

[12] A. Ziaei, A. Sangwan and J. H. L. Hansen, "Prof-Life-Log: Personal Interaction Analysis For Naturalistic Audio Streams," *IEEE Int'l Conference on Acoustics, Speech and Signal Proc. (ICASSP'13),* pp. 7770-7774, 2013.

[13] Apple Inc, "iPhone," *Web Page, Retrived [Sept. 2014] from Http://apple. Com/.,* 2014.

[14] Google Inc, "Google Glass," *Web Page, Retrieved [Sept. 2014] from Https://www. Google. com/glass/start/.,* 2014.

[15] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards and J. H. L. Hansen, "Signal Processing for Young Child Speech Language Development," *Workshop on Child, Computer and Interaction (WOCCI'08),* 2008.

[16] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proc. of the Nat'l Academy of Sciences of the USA,* vol. 107, pp. 13354-13359, Jul 27, 2010.

[17] H. Ali, A. P. Lobo and P. C. Loizou, "Design and evaluation of a personal digital assistant-based research platform for cochlear implants," *IEEE Trans. Biomed. Eng.,* vol. 60, pp. 3060-3073, Nov, 2013.

[18] O. Hazrati, S. O. Sadjadi and J. H. L. Hansen, "Robust and efficient environment detection for adaptive speech enhancement in cochlear implants," *IEEE Int'l Conference on Acoustics, Speech and Signal Proc. (ICASSP'14),* pp. 900-904, 2014.

[19] LENA Foundation, "LENA Research Foundation," *Webpage, Retrieved [Sept. 2014] from Www. Lenafoundation. Org/.,* 2014.

[20] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *Journal of the Acoustical Society of America,* vol. 124, pp. 3261-3275, 11, 2008.

[21] M. Garnier, N. Henrich and D. Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *Journal of Speech, Language and Hearing Research,* vol. 53, pp. 588-608, 2010.

[22] C. Yu, J. H. Hansen and D. W. Oard, "Houston, we have a solution': A case study of the analysis of astronaut speech during NASA apollo 11 for long-term speaker modeling," in *Fifteenth Annual Conference of the International Speech Communication Association,* 2014, .

[23] K. Sjölander and J. Beskow, "WaveSurfer-An Open Source Speech Tool," *International Conference on Spoken Language Processing (INTERSPEECH2000),* pp. 464-467, 2000.

[24] M. Brookes, "Voicebox: Speech processing toolbox for matlab," *Software, Available [Mar.2011] from Http://ee.Ic.Ac.uk/hp/staff/dmb/voicebox/voicebox.Html,* 1997.

[25] J. J. Godfrey, E. C. Holliman and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *IEEE Int'l Conference on Acou. , Speech, and Signal Proc. (ICASSP'92),* vol. 1, pp. 517-520, 1992.

[26] A. Sangwan, C. Yu, L. Kaushik and J. H. Hansen, "'Houston, we have a solution': Using NASA apollo program to advance speech and language processing technology." in *INTERSPEECH'13,* 2013, .

[27] A. Stolcke, "SRILM-an Extensible Language Modeling Toolkit," *Seventh International Conference on Spoken Language Processing (ICSLP'02),* 2002.