Impact of the Unknown Communication Channel on Automatic Speech Recognition: A Review

Jean-Claude Junqua

Panasonic Technologies Inc., Speech Technology Laboratory,
3888 State Street, Santa Barbara, CA, 93105, U.S.A.
E-mail: JCJ@Research.Panasonic.Com

I. ABSTRACT

This review article summarizes the main difficulties encountered in Automatic Speech Recognition (ASR) when the type of communication channel is not known. This problem is crucial for the development of successful applications in promising domains such as computer telephony and cars. The main technical problems encountered are due to the speaker and the task (e.g. speaking style, Lombard reflex, vocal tract geometry), the use of microphones with different characteristics, the variable quality of the support channels (e.g. telephone channels are noisy and have different characteristics), reverberation and echoes, the variable distance and direction to the microphone introduced by hands-free recognition, and the ambient noise which distorts the input speech signals. This overview characterizes and emphasizes these problems and highlights some promising directions for future research. Finally, it presents an attempt to characterize the sensitivity of a phoneme recognizer as a function of the source of channel distortion, using the TIMIT database and several of its variants (NTIMIT, CTIMIT, FFMTIMIT).

II. UNDERSTANDING CHANNEL-RELATED SOURCES OF VARIABILITIES/DISTORTIONS IN THE COMMUNICATION PROCESS

II.1 Introduction

Robust speech recognition deals with mismatches between training and testing [7]. Facing a wide range of unexpected adverse conditions, ASR systems need to be improved to cope with variabilities coming from the speaker, the type, direction, and position of the microphone, the transmission channel and the acoustic environment. These variabilities have been summarized and classified in three broad categories in Figure 1.

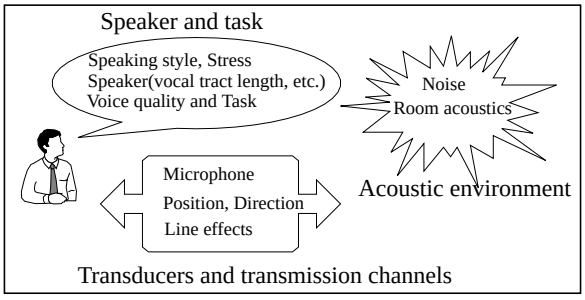

FIGURE 1. Main causes of variabilities which affect automatic speech recognizers.

II.2 Speaker Variability

Speaker-related variability is one of the main factors influencing current ASR systems. Voice quality, non-native speakers, stress-induced phenomena and speaker age groups are some examples of speaker-related acoustic variations. Among them, stress induced phenomena (e.g.: Lombard reflex [6]), age group differences (e.g. children versus adults [12]) and non-native speech productions) constitute real challenges for state-of-the-art ASR systems which, unfortunately, perform reasonably well only for carefully selected conditions.

II.3 Arbitrary Microphones

Microphones act as linear filters on the speech signal and account for different degrees of spectral slope depending on the microphone characteristics. This type of distortion is convolved with the speech signal. As the microphone-to-talker distance is often different, gain variation is also observed. Figure 2 provides a frequency-dependent Signal-to-Noise Ratio (SNR) and information about the spectrum tilt. To compute the SNR, hand-labels were used to distinguish noise from speech. The curves represent the average cepstrum of 370 words uttered by one male and one female speaker. Some important differences can be observed between the two microphones, especially in the low frequency range where the Sennheiser microphone has a higher SNR.

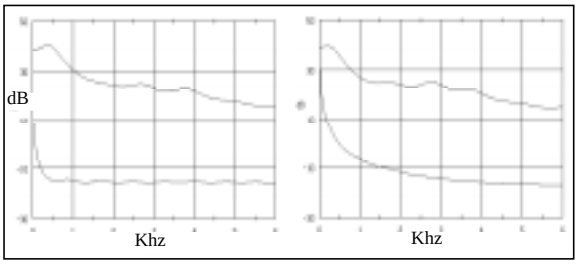

FIGURE 2. Average cepstrum obtained from a male and a female speaker recorded with a noise canceling HMD 414-6 Sennheiser (left figure) and a directional Ramsa WN-8080F (right) microphones. In each graph by taking a difference between the two curves a frequency-dependent SNR can be obtained.

Hands-free speech recognition is a challenging problem for which there are no completely satisfactory solutions. Interest in hands-free speech recognition came mainly from the use of speech over the telephone and in a car as well as in any situation where the talker's hands are devoted to another task.

In the case of hands-free speech recognition, the distance and the direction to the microphone vary. The speech signal is degraded because of echoes and ambient noise and the mismatch between training and testing is always variable. Figure 3 shows the frequency responses of the handset and speaker phone of a Panasonic telephone when the input was generated by a pulse generator through an artificial mouth at 89 dB SPL. We can see that there is a large difference between the two curves, espe-

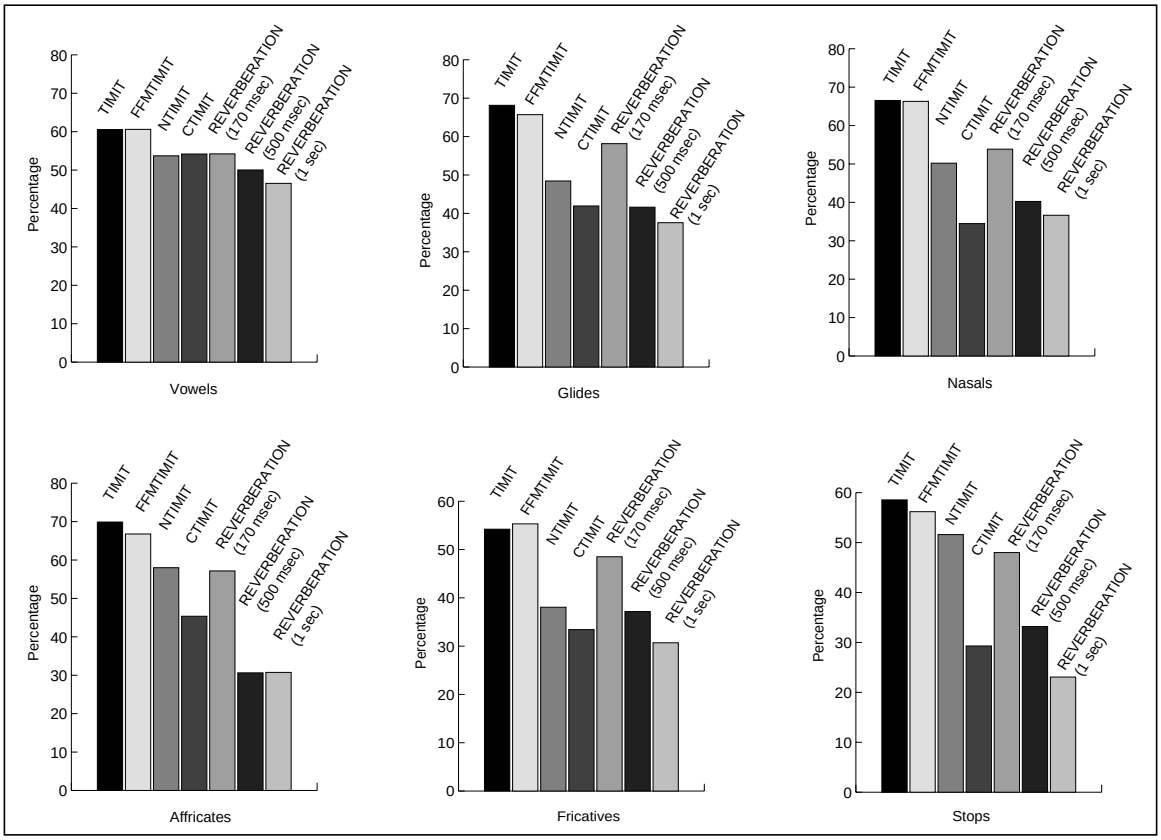

FIGURE 8. Phoneme recognition on broad phonetic classes for different databases.

tion of affricates more than the other classes of phonemes. In the case of telephone speech, fricatives seem to be the phoneme class that was most strongly affected. Finally, as the reverberation time increases, stops, fricatives and affricates get increasingly more difficult to recognize.

V. CONCLUSIONS

In this paper we described the main sources of distortions which affect the robustness of current ASR systems. Among them, speaker variability is probably the most challenging for ASR. After describing briefly each type of distortion we provided some new perspectives. To design ASR systems for multiple environment, adaptation should occur. On the fly speech adaptation, while still very difficult, is necessary.

There is also a need to conduct tests on real-world databases and to run controlled experiments to assess the robustness of our methods to one specific factor. Both types of tests/experiments are needed.

Finally, we should define standard tests and assess the sensitivity of our techniques across a wide range of distortions. In this paper we presented an attempt to do so for a phoneme recognizer evaluated on different variants of the TIMIT database representing various types of distortions. The TIMIT database and its variants can be a helpful resource to conduct such experiments. Recently, the HTIMIT database [9] was recorded to study handset transducer effects. This database adds another distortion dimension to the ones studied in this paper.

REFERENCES

[1] J.B. Allen and D.A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, pages 943–950, April 1979.

[2] P.J. Bloom. Evaluation of a dereverberation process by normal and impaired listeners. In *ICASSP*, pages 500–503, 1980.

[3] K.L. Brown and E.B. George. CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition. In *ICASSP*, pages 105–108, 1995.

[4] ESCA-NATO. *Proceedings of Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, France, J-C. Junqua and J-P. Haton, editors*. 1997. April.

[5] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *ICASSP*, pages 109–112, 1990.

[6] J-C. Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20:13–22, November 1996.

[7] J-C. Junqua and J-P. Haton. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers, 1996.

[8] K.-F. Lee. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. PhD thesis, 1988. Carnegie Mellon University.

[9] D. Reynolds. HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects. In *ICASSP*, pages 1535–1538, 1997.

[10] D. Reynolds, M. Zissman, T. Quatieri, G. O'Leary, and B. Carlson. The effects of telephone transmission degradations on speaker recognition performance. In *ICASSP*, pages 329–332, 1995.

[11] E. Thelen, X. Aubert, and P. Beyerlein. Speaker adaptation in the Philips system for large vocabulary continuous speech recognition. In *ICASSP*, pages 1035–1038, 1997.

[12] J. Wilpon and C. Jacobsen. A study of speech recognition for children and the elderly. In *ICASSP*, pages 349–352, 1996.

III. NEW PERSPECTIVES FOR ROBUST ASR

Recent techniques for robust speech recognition focused mainly on 1) robust signal pre-processing techniques and 2) feature and model compensation [7]. The method selected has to do with the space in which the mismatch is dealt with: signal, feature or model. Among the methods developed to handle the mismatches between training and testing data, adaptation/compensation techniques are getting much interest because of their ability to handle a large range of channel and noise variations together with speaker and speaking style differences. Recently (e.g. [11]) the combination of different adaptation techniques inside the same system was shown to provide improved performance. However, while a human is able to adapt to a new voice with a minimal amount of training data, unsupervised instantaneous adaptation is still a major challenge to machines. The emerging interest for adaptation as a technique to solve the robustness problems comes from the good level of performance now reached by ASR systems and our effort to address real-world applications.

While adaptation techniques provide a useful path to improve the robustness of our systems and to address a wide range of issues, several additional directions for improvement were discussed at the ESCA-NATO workshop "*Robust speech recognition for unknown communication channels*" in Pont-à-Mousson, France [4]. These include:

- the use of partial information. With such methods unreliable sources of information are ignored (e.g. low SNR regions);
- the independent processing and recombination of several feature streams;
- the use of several representational dimensions and time window lengths to expand our models and capture different and complementary sources of information;
- the importance of using a universal framework which can cope with all the distortions rather than a framework which operates in a specific controlled condition;
- the necessity to develop a standard set of tests with a wide range of distortions (e.g. reverberation, telephone line effects, noise) to evaluate our new approaches and characterize the sensitivity of our methods to the source of distortion.

These directions, together with on-line fast adaptation constitute new perspectives for robust speech recognition. In the next section we will present an attempt to characterize the sensitivity of a phoneme-based recognizer to the source of distortion. Such a study is useful to provide insights on the strengths and weaknesses of the techniques evaluated and errors made by ASR systems.

IV. CHARACTERIZATION OF THE SENSITIVITY OF A PHONEME RECOGNIZER TO THE SOURCE OF DISTORTION

IV.1 The Experiments

We evaluated an HMM continuous density phoneme recognizer trained on several variants of the TIMIT database. 12 MFCC coefficients combined with the normalized log energy and first order regression coefficients (26 coefficients) were used in the speech parametric representation. Long-term cepstral mean normalization was applied to the cepstral vectors. 4 Gaussian densities per state and 3 state HMMs were trained on the SI and SX training sentences of the training database. The test part of the TIMIT and related databases were used for testing. 48 phonemes were used for the computation and 39 phonemes (as defined in [8]) for computing the results. In all tests, the insertion rate was kept around 10% and the percentage of correct phones was evaluated.

We assessed the effect of the bandwidth limitation, telephone line effects (NTIMIT), cellular network distortions (CTIMIT), the use of a far field microphone (FFMTIMIT) and simulated reverberation. Different reverberation times were simulated with a reverberation simulation program [1]. The effect of each distortion was separately evaluated and the errors on broad phonetic classes were analyzed.

IV.2 Results

Figure 6 shows the effect of bandwidth limitation on the TIMIT and FFMTIMIT databases. As expected, limiting the bandwidth results in a slight decrease in performance (between 2 and 3%). While results on FFMTIMIT are slightly less accurate than on the TIMIT database, performance does not degrade much. This is probably due to the high quality of the FMMTIMIT database which does not represent what we may expect with real-conditions using far field microphone recording.

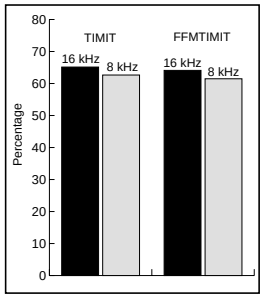

FIGURE 6. Effect of downsampling on the TIMIT and FFMTIMIT databases (the Y axis indicates the % of phonemes correct).

Figure 7 shows comparative performance between the different conditions evaluated. It also presents the effect of the reverberation time on recognition.

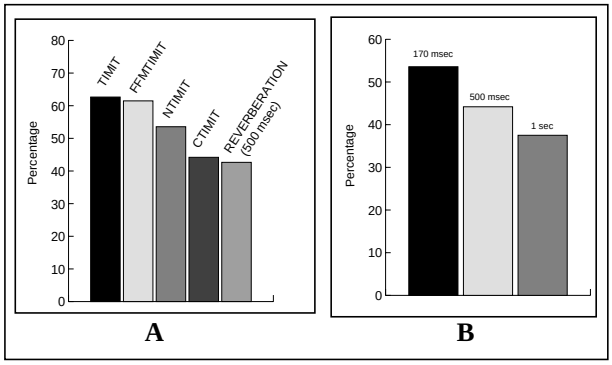

FIGURE 7. Comparative performance of phoneme recognition on (A) different databases representing various channel distortions and (B) with different reverberation times.

It can be seen that reverberation and distortions due to the cellular network are the most difficult to deal with. Furthermore, phoneme recognition performance decreases as reverberation time increases.

By looking at Figure 8 it can be seen that recognition performance on CTIMIT is not very effective on consonants, while long-term reverberation affects the recogni-

cially in the low and high frequency ranges. These differences, which depend on the distance and direction of the speaker to the microphone and the fact that ambient noise will be captured easily, are the main sources of difficulties for hands-free speech recognition.

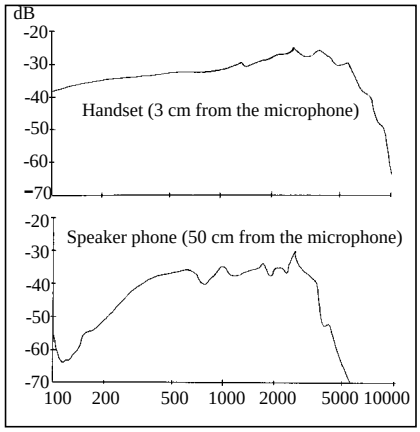

FIGURE 3. Frequency responses of the handset and speaker phone of a Panasonic telephone when the input was generated by a pulse generator through an artificial mouth at 89 dB SPL.

II.4 Telephone Channels

Telephone speech is more difficult to recognize than high quality speech due to bandwidth limitation, handset and connection quality variations, and increased background noise. Callers use speaker phones, cellular phones, and ordinary telephones with varying microphone and transmission quality. The speech comes from an uncontrolled environment containing different types of background noise, such as television, radio, and other speech. The main sources of channel distortion over telephone lines can be separated in various categories such as burst or impulse noise, hum, additive stationary noise, inter-modulation distortion, echo, frequency translation, unknown channel gain and phase response, added low-frequency tones, and breath intake and release. Some of these distortions are additive in the spectral domain and others are additive in the log spectral or cepstral domain (convolutional distortions). Figure 4 shows three spectrograms of the same sentence from the TIMIT, NTIMIT [5] and CTIMIT [3] databases.

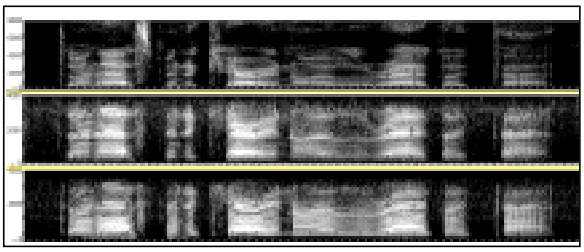

FIGURE 4. Spectrograms of the sentence “Don’t ask me to carry an oily rag like that” from respectively the TIMIT (top), CTIMIT (middle) and NTIMIT (bottom) databases.

The TIMIT sentence was recorded under almost ideal conditions, while in NTIMIT the sentence was played through a carbon-button telephone handset and transmitted through the telephone network. In the case of CTIMIT, the TIMIT sentence was transmitted over the cellular network. It can be noted that besides some noise addition due to the telephone and cellular networks, the CTIMIT and NTIMIT sentences show evidences of some non-linear effects. In [10] it was pointed out that the carbon-button microphone used to record NTIMIT produced some nonlinear distortion introducing “phantom formants” and some changes in the formant bandwidths. This spectral shaping due to the type of microphone accounts for some performance loss when the same microphone is not used during the training phase.

II.5 Reverberation

Figure 5 shows the impulse response obtained with a room reverberation simulation program presented in [1]. This simulation program models the effects of echo and reverberation encountered in an enclosure with sound-reflecting walls. In Figure 5 the reverberation time is less than 550 ms.

Reverberation degrades speech intelligibility to a large extent through the masking of direct sounds by reflected energy. Intelligibility seems to be affected not so much by the early (<30 ms) pattern of reflections (color) but by the reverberant tail generated when one speech sound arrives sufficiently late to overlap the time-waveform of a later, direct speech sound [2]. In the frequency domain, the direct speech energy is masked by noise with a speech-shaped spectrum. Therefore the positions in time and the order of the speech components influence the type of masking.

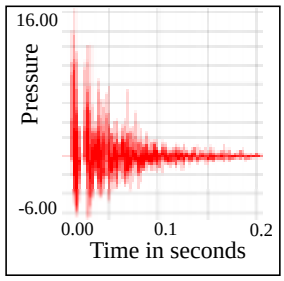

FIGURE 5. Impulse response for a room size (in feet) of 10’ x 11’ x 12’, with all wall reflection coefficients being equal to 0.9, the talker coordinates at (9’, 8’, 11’) and the microphone coordinates at (1’, 1’, 2’).

II.6 The Acoustic Environment

In ASR, performance is rather uniform for SNRs greater than 25 dB but there is a very steep degradation as the noise level increases. Although much effort has been made to improve the robustness of current recognizers against noise, many algorithms still assume low noise or model the masking noise with stationary white-Gaussian or pink noise which does not always represent realistic conditions. Collecting data in an operational environment is a key factor to understanding and solving real problems. However, real-world speech databases necessitate great efforts to produce meaningful task scenarios and are very expensive to develop.

When speech is produced in noise there is a modification of speech production leading to the Lombard reflex. To simulate this reflex, several databases were recorded while the speakers listened to noise through headphones. A great variability in the increase of vocal effort was observed. This probably comes from the way databases are recorded. Generally, it is implicitely assumed that the Lombard reflex is a physiological effect. However, it seems that in the real world, the magnitude of the response of the speakers is governed by the desire to obtain intelligible communication [6]. To better understand the Lombard reflex, more realistic databases need to be recorded.