Ubiquitous Speech Communication Interface

B.H. Juang
AVAYA Labs Research, Basking Ridge, New Jersey, USA
bjuang@avaya.com

Abstract

The Holy Grail of telecommunication is to bring people thousands miles apart, anytime, anywhere, together to communicate as if they were having a face-to-face conversation in a ubiquitous tele-presence scenario. One key component necessary to reach this Holy Grail is the technology that supports hands-free speech communication.

Hands-free telecommunication (both telephony and teleconferencing) refers to a communication mode in which the participants interact with each other over a communication network, without having to wear or hold any special device. For speech communications, we normally need a loudspeaker, a microphone or a headset. The goal of hands-free speech communication is thus to provide the users with an intelligent voice interface, which provides high quality communication and is safe, convenient, and natural to use. This goal stipulates many challenging technical issues, such as multiple sound sources, echo and reverberation in the room, and natural human-machine interaction, the resolution of which needs to be integrated into a working system before the benefit of hands-free telecommunication can be realized. In this paper, we analyze these issues and review progress made in the last two decades, particularly from the viewpoint of signal acquisition, restoration and enhancement. We lay out new technical dimensions that may lead to further advances towards realization of a truly ubiquitous speech communication interface to an intelligent information source, be it a human or a machine.

** Part of this paper has been published in Proceedings of Workshop on Hands-free Speech Communications 2001, ATR, Kyoto, Japan, April 2001.*

1. Ubiquitous Speech

On March 10, 1876, Alexander Graham Bell uttered, "Mr. Watson, come here, I want to see you," into his speech transmission implement, which later became an indicator of civilization and a household name called *telephone*. He wrote to his father that evening, "I feel that I have at last found the solution of a great problem, and the day is coming when telegraph wires will be laid on to houses just like water or gas is, and friends will converse with each other without leaving homes." That marks the birth of modern telecommunication, particularly in the sense of speech communications. Alexander Graham Bell’s invention also inspired people’s imagination towards the Holy Grail of telecommunication, which is to bring people thousands miles apart, anytime, anywhere, together *to communicate as if they were having a face-to-face conversation*. We thus aspire to go from telephony to ubiquitous tele-presence.

In order to achieve this Holy Grail, one key component necessary to enable tele-presence is the technology that supports hands-free, ubiquitous speech communication.

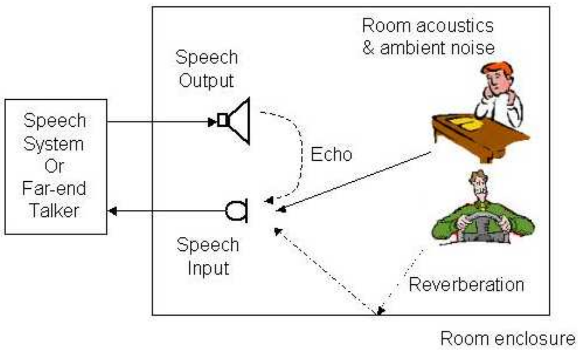

Figure 1 Scenario of hands-free communications

Hands-free telecommunication (both telephony and teleconferencing) refers to a communication mode, in which the participants speak and interact with each other over a communication network, without having to wear or hold any special device such as a microphone or a headset. Fig. 1 depicts the scenario of hands-free telecommunication, taking place in an enclosed room, such as an office or the cabin of a car. The room (i.e., the near-end) is equipped with a transducer assembly (a microphone or an array of microphones), which picks up the acoustic signal in the room, and a loudspeaker, which plays out the signal from the remote end. In such a communication configuration, the talker in the room is usually situated at a certain comfortable distance from either the microphone assembly or the loudspeaker, or moving around the room without holding or wearing any additional device. This communication scenario is quite different from traditional telephony in which a telephone with a handset is regularly used. A relevant device called Speakerphone that exists today may allow a primitive and limited form of hands-free telecommunication. Note that at the remote end (or the far-end), it may be another human or a machine that attempts to communicate with the user at the near-end.

There are a number of strong motivations behind ubiquitous speech and hands-free telecommunication. First, people want mobility, even just locally. Communication using a tethered device is both inconvenient and undesirable. The use of a Speakerphone is exactly to let the user do away with a locally tethered device. Second, the idea of wearing a wireless device, such as an FM headset, has not become popular among users due to a number of non-technical reasons. Third, in mobile communications, the concern over safety is growing; many municipalities and countries in the world are erecting legislation to disallow a driver to use a hand-held cellular phone while driving. A hands-free communication device installed in the car would alleviate the distraction of a hand-held cellular phone. Fourth, people are constantly looking for improved communication quality and naturalness in interface. A speakerphone that uses gain switching to provide limited full-duplex conversation cannot deliver high speech quality to allow proper use of a remote speech recognizer or to support multi-point teleconferencing without causing frustration on the participants. Many signal processing problems need to be solved in order to be able to realize a high-quality hands-free telecommunication system.

2. A Complex Signal Processing System

A number of critical technical issues are involved in the hands-free telecommunication paradigm. The acoustic signal picked up by the microphone includes that of the far-end played out from the loudspeaker (which we refer to as "echo" or "system echo" to distinguish it from reverberation below), the near-end talker's speech, and a substantial amount of the near-end ambient noise. Unlike a handset, which responds primarily to the talker holding the device, the microphone used in hands-free communications receives a multiplicity of sound sources. For most of the current automatic speech recognition systems designed to respond to a single talker's speech, this is a major source of degradation in recognition performance. In addition, since the talker and the microphone are not expected to stay at a fixed relative position, the sound quality in the system will certainly vary. The acoustic signal entering the system is a strong function of the room as well as the type of microphone that the system uses. The room inflicts two essential effects on the acoustic signal, one called colorization and the other reverberation. Colorization refers to the change of short-time spectral shape the room causes on the source signal. The effective length of the impulse response of a room is a function of the room configuration (geometric shape of the room and acoustic reflectivity of the wall). When the impulse response is more than a few tenths of a second long, the reverberation effect becomes noticeably disturbing. Both a human listener and a speech recognizer would react negatively to reverberation. Another important issue is the support of natural communication interaction, such as "barge-in" in speech recognition. With hands-free communication, participants expect and are expected to behave as if a face-to-face conversation is taking place. Thus, interruptions, barge-ins, and even sidebars would occur more frequently than when using a telephone handset. Technical implications due to these behaviors are the need for an improved speech activity detector, and a reliable echo canceler to allow full-duplex communication and proper barge-in for natural human-machine interaction. With the fragile performance of today's speech recognizer, integration of an echo canceler into the system for a reliable, natural interaction with the machine is still an interesting challenge. These issues, namely, multiple sound sources, echo and reverberation, and naturalness in interactive behaviors, make hands-free telecommunication one of the most intriguing signal processing problems in modern days.

The purpose of this paper is to review the progress towards hands-free telecommunication in the past two decades, particularly from the viewpoint of signal restoration and enhancement, and to draw new technical dimensions, which may stimulate additional advances. These include: the noise problem, the duplex problem, the colorization problem and the reverberation problem. To simplify and systemize the approach to the challenge, we begin with a taxonomical view of the problems involved. For each problem area, we highlight the significance of key advances and point out new technical directions that deserve further close investigation.

3. The Noise Problem

We consider all acoustic signals received at the microphone other than the intended or authorized talker to be noise. It thus includes ambient noise, machine noise (e.g., a personal computer), extraneous speech such as background conversation, and competing speech (voice from other participants of the session), Noise is normally assumed to be additive except in rare situations.

There are many causes, such as saturation, clipping, or level fluctuation, of perceptual detriment to a speech signal and additive noise is perhaps the most pervasive one among all. Human perception of sound quality is a complex phenomenon and there is evidence that subjective judgment on sound quality may be quite different when it comes to additive (e.g., noise) and convolutive (e.g., filtering) modifications. Subjective quality is often measured by MOS, the Mean Opinion Score, which is based on a listener's judgment of the quality in a scale from 1 (bad) to 5 (excellent). Figure 2 is a plot of the average MOS over a population of listeners as a function of the MNRU (modulated noise reference unit) in dB. MNRU is a type of signal to noise ratio (SNR). Shown in the figure are two curves, pertaining to two different narrowband source materials, with and without being subject to IRS filtering, respectively [1]. It can be seen that the subjective quality is a monotonic function of the MNRU SNR, dropping rapidly as SNR decreases below 30 dB SNR. However, what is interesting is that the subjective quality judgment for the filtered speech is in fact higher than that of the unfiltered one. Filtering which in general introduces distortion to the source material can actually improve the perceived quality in a particular setup. This shows that signal restoration and enhancement methods, while may improve upon the subjective quality of the speech material, may and very likely will introduce objective distortion to the signal. One thus needs to take particular note in setting the goal of signal enhancement for hands-free communication.

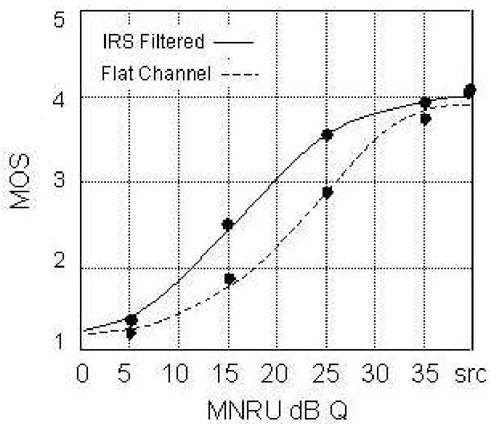

Figure 2 Perceptual judgment on the effect of noise and distortion upon the speech signal

Noise cancellation is one of the early ideas to deal with the problem of additive noise. Fig. 3 illustrates the fundamental concept in noise cancellation. It utilizes two input signals, usually one being the noisy signal and the other the reference noise. A filtered version of the reference noise is used to cancel the noise that contaminates the signal. The filter is optimized such that the output signal power is minimized. In order for the method to work, there must exist sufficient coherence between the reference noise and the noise component in the noisy speech. This situation is somewhat contrived because in most applications such as hands-free communication, it is impractical to assume the availability of such a reference noise signal. Most often, the two channels of input would contain a mixture of the desired speech signal and the noise, to a varying degree, respectively. For applications in cars, studies show that in order for the noise components to be sufficiently coherent, the two microphones cannot be more than 50cm apart [2]. These conflicting factors make this early approach impractical for the application considered here.

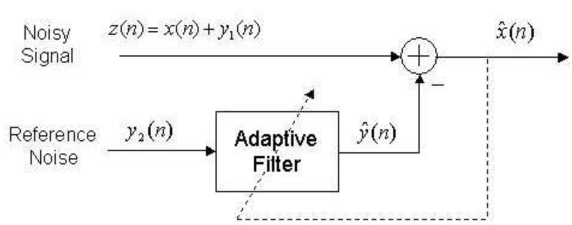

Figure 3 Noise cancellation by adaptive filtering methods

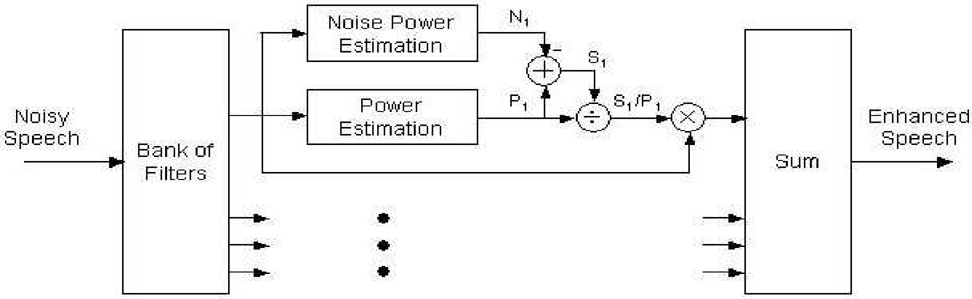

Figure 4 Signal enhancement based on spectral subtraction and Wiener Filtering

Methods based on short-time power spectral estimates developed over the last two decades proved to be somewhat effective. This class of noise suppression methods, a conceptual diagram of which is shown in Figure 4, includes spectral subtraction [3], Wiener filtering [4] and various extensions thereof. In essence, the method estimates the noise power spectrum, often over a period when there is no speech, subtracts it from the power spectrum of the noisy signal, and then re-synthesizes the "enhanced" signal using the noisy phase information. For a stationary signal process contaminated by a stationary noise process, the optimal linear estimator that achieves minimum mean square error is the so-called Wiener filter, characteristics of which is defined by $P_s/(P_s+P_n)$ where P_s and P_n are the power spectral density of the signal and the noise, respectively. Realization of the Wiener filter usually involves approximation to alleviate the issue of non-causality and the power spectral density function is replaced by a number of ad hoc power spectral estimates, which are non-trivial and difficult to analyze in the case of noisy speech due to its non-stationarity. Despite the discrepancy between the theory and its realization, this class of methods is being used widely in many applications with a certain degree of success. In general, the enhanced speech sounds "quieter" or "less noisy" but also comes with undesirable artifacts, which may be detrimental to the performance of speech recognizers.

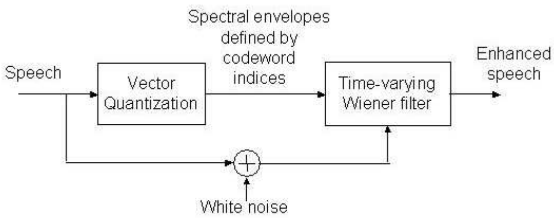

Figure 5 Setup of a noise suppression experiment based on time-varying Wiener filtering

Since approximation is necessary in applying Wiener filtering to the problem of noise suppression, a question thus arises on the relationship between the approximation and the perceived quality in the enhanced signal. To gain some insight into the issue, an experiment as depicted in Fig. 5 was conducted. A "clean" speech sample was analyzed and its resultant 10-th order all-pole model spectral sequence was then vector-quantized using the likelihood ratio distortion measure. A sample of white noise was then digitally added to the speech material to create an artificial noisy speech signal. The sequence of vector-quantized spectral envelopes (as defined by the all-pole models) was then used as an estimate of the short-time speech power spectra, which were combined with the estimated noise power spectrum to form a sequence of "short-time Wiener filters". The short-time Wiener filter sequence was then used to filter the noisy speech. Vector quantizers of various resolutions, ranging from 32 codewords to 128 codewords, were tested to compare the enhanced results. It was noted that the spectral distortion was measured at 3.8, 2.95, and 2.69 dB, respectively, for 32-, 64- and 128-codeword vector quantizer cases. Informal listening showed that a resolution of 64 codewords would be sufficient to virtually eliminate the perceived noise in the speech sample. There is a number of implications in this result. First, the error in power spectral estimation for the clean signal for use in Wiener filtering can be as much as 3 dB without perceptually noticeable degradation. Second, the problem of power spectral estimation is being transformed into a detection problem – determining which power spectral model in the rather small set of codewords for use in the Wiener filter. This prompted the idea of a maximum posteriori probability (MAP) approach to the enhancement problem [5] as shown in Fig. 6. The MAP method aims at maximizing the joint probability of the signal and the noise given the speech and noise models, trained separately as the prior knowledge of the signal sources. The method achieves the optimization objective through an iterative maximization procedure and was shown to be able to reduce the noise perception substantially. The MAP method also provides a framework for introducing source knowledge, which is embedded in the speech model, represented either by the vector quantization codebook or a hidden Markov model. The concept of source embedding differentiates itself from other ad hoc power spectrum smoothing schemes, commonly used in the traditional spectral subtraction and Wiener filtering method, and represents a recent advance in noise suppression.

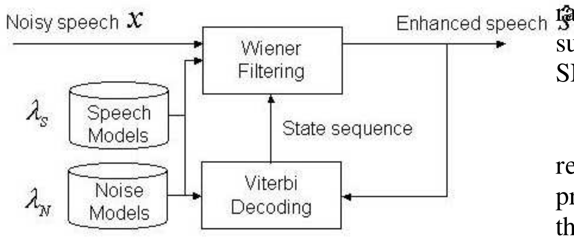

Figure 6 An iterative Maximum A Posteriori approach to signal enhancement using decision-feedback

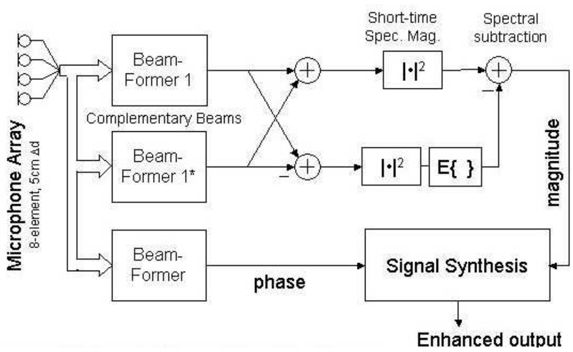

Figure 7 Spectral subtraction using complementary beamforming

Integration of the short-time power spectrum methods with multiple transducers is another significant advance in speech enhancement. Multi-transducer signals can be obtained by use of a traditional delay-sum array [6], which has a directivity gain depending on the configuration of the array (including the number of transducer elements used and their geometry). The directivity of an array provides a substantial SNR improvement when the source and the noise are coming from different angles relative to the beam pattern of the array. Spectral subtraction method can be applied to the signal for further removal of the residual noise effect. The work of Saruwatari et. al. [6], depicted in Fig. 7, represents yet another major contribution, which attempts to integrate multi-transducer signals with spectral subtraction. A pair of complementary beams are formed to obtain improved power spectrum estimate of the signal as well as that of the noise A properly designed microphone array offers an SNR improvement in the range of 15-20 dB when the noise is more than 30^0 off the center or look direction. (The difference between a traditional delay-sum array and the one using complementary beams is noticeable but somewhat moderate, in the range of less than 1 dB.) Coupling with spectral subtraction, an additional 2-3 dB improvement in SNR is obtainable.

The use of multi-transducer inputs has in recent years transformed the noise suppression problem into the problem of source separation. In the formulation as depicted in Fig. 8, the received signal is assumed to be the result of mixing and the goal of signal separation is to de-mix the multi-channel signals such that each channel would produce a single source output. To solve such as problem, one usually assumes that the signals from various sources are uncorrelated, and the design of the de-mixer is to de-correlate the output signals. While other constraints and optimization objectives are possible, current work focuses more along the classical concept of matrix diagonalization and principle component analysis. Some preliminary results obtained under benign conditions, such as short room impulse response and a quiet room, are promising.

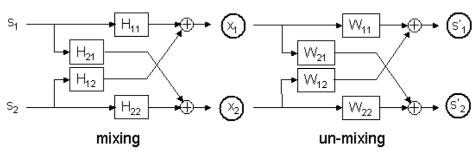

Figure 8 Mixing and un-mixing of multi-channel signals

4. The Duplex Problem

A full-duplex communication system allows the users on both ends of the connection to send and receive signal at any time during the session in an unregulated manner. In traditional telephony, this is accomplished by either a 4-wire connection (one pair for transmit and the other for receive) or a hybrid coil in the center office that connects a 2-

wire local loop equipment to a 4-wire trunk, using a network echo canceler on the 4-wire side to cancel the far-end signal. In hands-free telecommunication, an acoustic echo canceler is needed to suppress the far-end signal that is being played out through the loudspeaker at the near-end. While effective echo cancellation algorithms are well known, their application to hands-free communication deserves some additional attention.

An echo canceler works by estimating the impulse response of the echo return path, which is then used to convolve with the far-end signal to produce an estimated echo signal for cancellation purposes. Estimation of the impulse response of the acoustic echo return path is one of the main challenges here. The length of the room impulse response varies substantially according to the room size and shape, and the wall material. The room impulse response is also a rather sensitive function of the room temperature, although our auditory system does not seem to display such. With the participants moving freely in the room, the echo canceler needs a rapid impulse response tracking ability in order to cope with the changing conditions. Compared to network connection, characteristics of which does not vary substantially once the connection is made, the need in tracking the room impulse response in dealing with the acoustic echo is much more imperative.

Several algorithms with fast convergence properties were investigated in the past few years. Progress was obtained in frequency domain adaptive echo cancellation [7], which takes advantage of the property of a circulant matrix for efficient approximation to the solution process. Another technique capitalizing on the block least squares algorithm [8] was implemented specifically as an hands-free communication interface with an automatic speech recognition system, to enable a smooth and reliable barge-in interaction with the machine.

In hands-free communications, the problem of double-talk is also more severe and harder to deal with. A double-talk situation arises when talkers on both ends of the connection speak. During double-talk, the echo canceler stops adapting the filter coefficients to prevent system divergence. In a hands-free conferencing environment, there will be more talkers speaking or barging in at the same time, as any natural human-to-human interaction would be. Integration of a double- or multiple-talk (when more than two parties are involved) detector with the echo canceler at each end is undoubtedly going to be a major system challenge, not to mention the detection algorithm itself.

Most of double-talk detection algorithms [9] were developed for network applications. When applied to hands-free communications, there is a need to revisit these algorithms for optimal system design.

5. The Colorization Problem

The colorization problem refers to the change in the short-time spectral shape of the source signal, when played out through the loudspeakers, due to the room impulse response. The colorization problem has more to do with the performance of the acoustic echo canceler than with the human auditory perception. In the current context, the colorization problem is thus considered part of the echo cancellation problem.

6. The Reverberation Problem

When the impulse response of the room is long, it causes reverberation. Reverberation has a detrimental effect both on perception and on echo control for hands-free communications. When the reverberation time is a few times longer than what is often referred to as the syllabic interval, the intelligibility of the signal in the room is seriously compromised, making it difficult to conduct voice conversation. For automatic speech recognition, the usual techniques such as cepstral subtraction would have rather limited use in dealing with such a situation, as shown in Fig. 9.

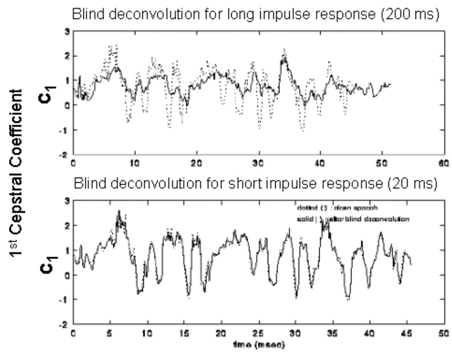

Figure 9 Effects of reverberation on speech analysis

The two panels in Fig.9 show the effect of blind deconvolution by cepstral subtraction on contours of the first cepstral coefficient (c_1) of a sentence, after the signal has been convolved with a long impulse response (200 ms, upper panel) and a short impulse response (20 ms, lower panel), respectively. The dotted line represents the cepstral contour of the original clean signal. As can be seen, the blind deconvolution technique was able to recover most of the distortion introduced by reverberation when the impulse response is short. When the impulse response is long, it is not effective at all.

A formal approach to dealing with reverberation is by inverse processing, which involves estimation of the (long) room impulse response and inverse filtering, as depicted in Fig.10. However, it is well known that a room impulse response is usually non-minimum phase and its corresponding inverse filter is thus unstable. Furthermore, since the impulse response is also long, the computational complexity in both estimation and inversion is usually prohibitive. The MINT theorem [10], which proves the existence of an exact solution when two microphones are used, provides a theoretical framework for multi-channel solutions. However, most practical solutions rely on error minimization.

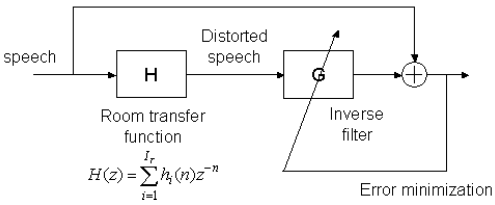

Figure 10 De-reverberation by inverse processing

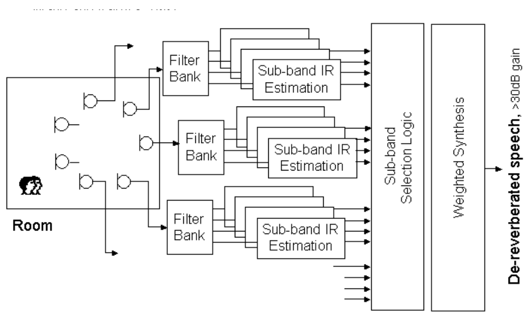

Figure 11 De-reverberation by composite inversion using multiple transducers

Several interesting new ideas have been proposed for estimation and inversion of the room impulse response. The work by Wang and Itakura in 1991 [11], illustrated in Fig. 11, proposes to use multi-transducer and multi-band analysis, from which a room impulse response within each band from each transducer is estimated. A selection-logic is then applied to choose a proper set of such partial impulse responses for final synthesis of the de-reverberated signal. A de-reverberation gain of over 30 dB has been reported, although the algorithm is sensitive to the room condition such as the temperature, which may cause significant variation on the impulse response.

In another study [12], a successive inversion algorithm, also utilizing a multi-transducer setup, was investigated.

7. The System Problem

As ubiquitous and hands-free speech interface technologies become more mature, the system becomes less visible to the user. And as the user becomes less conscious about the need to work with the system, the technological stress imposed on the system will inevitably increase. Careful system design that takes into account human behaviors in communication interactions will then also become imperative.

8. Summary

Ubiquitous speech interface enables people to communicate naturally with each other for optimal sharing of information or with a machine for convenient access to information. To afford a ubiquitous speech interface, the technology of hands-free communication is essential. Hands-free communications, however, entail a wide range of challenges, including the noise problem, the full-duplex and echo control problem, the colorization problem, and the reverberation problem. Signal processing techniques that aim at restoration and enhancement of the signal are at the core of these challenges. Furthermore, as research progress continues in the basic signal processing area, there is an increasing need to understand the issue of system integration and operation, which demands a rigorous study on the human behavior in order for the machine to support natural human interface.

Reference

1. Peter Kroon, “Evaluation of Speech Coders,” Chapter 13 in *Speech Coding and Synthesis*, Klein and Paliwal (ed.), Elsevier, 1995.

2. N. Dal Degan and C. Prati, “Acoustic noise analysis and speech enhancement techniques for mobile radio applications,” *Signal Processing*, 15, pp. 43-56, 1988.

3. S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust. Speech & Signal Processing,* ASSP-27, No.2, pp.113-120, April 1979.

4. Jae S. Lim and Alan Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp.1586-1604, December 1979.

5. Y. Ephraim, D. Malah and B. Juang, “On the application of hidden Markov modelos for enhancing noisy speech,” *IEEE Trans. Acoust., Speech & Signal Processing*, ASSP-37, 12, pp.1846-1856, December 1989.

6. H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, ``Speech enhancement using nonlinear microphone array based on complementary beamforming," IEICE Trans. Fundamentals, vol. E82-A, no.8, pp.1501--1510 (1999)

7. J. Benesty and D.R. Morgan, “Frequency domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation,” ICASSP-2000, Istanbul, June 2000.

8. E. Woudenberg, F.K. Soong and B.H. Juang, “A block least squares approach to acoustic echo cancellation”, ICASSP-99, Phoenix, March 1999.

9. T. Gaensler, J. Benesty and S.L. Gay, “Double-talk detection schemes for acoustic echo cancellation,” in *Acoustic Signal Processing for Telecommunication*, Gay & Benesty (ed.), p. 81-100, Kluwer, Boston, 2000.

10. M. Miyoshi and Y. Kaneda, “Inverse Filtering of Room Acoustics,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-36, No. 2, p. 145-152, 1988.

11. H. Wang and F. Itakura, “An approach to dereverberation using multi-microphone sub-band envelope estimation,” ICASSP-91, 953-956, vol.2H. , 1991.

12. Alice Wang, Frank Soong and B.H. Juang, unpublished technical report, 1997.