論文 / 著書情報
Article / Book Information

Title	Toward Robust Speech Recognition Under Adverse Conditions
Authors	Sadaoki Furui
Citation	Proc. ESCA Wrokshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu, , , pp. 1-12
Pub. date	1992, 11

ESCA Workshop on Speech Processing in Adverse Conditions
Cannes-Mandelieu, France, November 10-13, 1992

Toward Robust Speech Recognition Under Adverse Conditions

Sadaoki Furui

NTT Human Interface Laboratories, 3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

Speech signals are subject to variations resulting from linguistic variables as well as such acoustic variables as additive noise, speaker individuality, environment-dependent speaking styles, and microphone and transmission characteristics. This paper overviews the main methods that have been investigated to cope with these variations, which are the major factors degrading the performance of speech recognition systems used in practical situations.

The following methods have been used to deal with additive noises: using special microphones, using auditory models for speech analysis and feature extraction, reducing and suppressing noise, using noise masking and adaptive models, using spectral distance measures that are robust against noises, and compensating for spectral deviation resulting from the special speaking manners used in noisy environments (Lombard effect). Various methods have also been used to cope with the problems caused by the different characteristics of different kinds of microphones.

Discourse recognition using spontaneous speech has recently occupied the attention of many researchers. In this area, it is necessary to cope with variations that are not encountered when recognizing speech read from a text. Various approaches to giving recognition systems the ability to automatically adapt to individual speakers have also been actively explored. To cope with the variation related to linguistic processing, methods of adaptation to a new task have been investigated.

1. Introduction

A panel session titled "Towards robust speech recognition" was organized in the IEEE Speech Recognition Workshop held at the Arden House, New York in December 1991. This session addressed the problem that even if a speech recognition system performs remarkably well in laboratory evaluations and during demonstrations to prospective clients, it often performs not nearly as well in the "real world". This is mainly because the speech that actually has to be recognized varies for many reasons and therefore usually differs from training speech. Even the psychological awareness of communicating with a speech recognizer could make the talker produce a noticeable difference.

The recognition performance for speech influenced by noise and distortion is especially degraded when only very clean speech was used to train the system. The main causes of speech variation can be classified according to whether they originate in the speaking and recording environment, the speakers themselves, or the input equipment (Table 1).

Although the physical phenomena of variation can be classified as noise addition or as distortion, the distinction between these categories is not clear. The actual phenomena resulting from variation are the following:

Table 1 - Main causes of speech variation.

Environment	Speech-correlated noise reverberation, reflection Uncorrelated noise additive noise (stationary, unstationary)
Speaker	Attributes of speakers dialect, gender, age
	Manner of speaking breath & lip noise stress Lombard effect rate level pitch cooperativeness
Input equipment	Microphone (Transmitter) Distance to the microphone Filter Transmision system distortion, noise, echo Recording equipment

(1) Spectral variation

Several spectral peaks disappear and the spectral envelope becomes flat. Spurious spectral peaks appear. Spectral inclination and bandwidth change. Other spectral variations include nonlinear transformation.

(2) Nonlinear time expansion and contraction

(3) Additive noise increases the difficulty of determining the speech period.

When people speak in a noisy environment, not only does the loudness (energy) of their speech increase, but the pitch and frequency components also change. These speech variations are called the Lombard effect: the first formant of a vowel often increases while the second formant decreases. The Lombard effect also includes a significant change in spectral tilt [1-3]. Several experimental studies have indicated that these indirect influences of noise have a greater effect on speech recognition than does the direct influence of noises entering microphones [4-5]. It has also been reported that articulation effects under noisy conditions are context dependent [3].

Because methods for coping with individual speech variation have already been surveyed [6-7], this paper just briefly reviews the major speaker-independent methods used in speech recognition, concentrating especially on speaker adaptation techniques. Readers who want to know details should refer to the original papers.

When recognizing spontaneous speech in dialogs, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, style shifts, and instability in the detection of end-points.

Another problem related to the variation of speech is how to adapt to a new task. Most advanced speech recognition systems use statistical approaches consistently for acoustic and linguistic processing. To build reliable statistical models, it is necessary to use large speech and text databases. These databases must be changed to fit new tasks every time the tasks are changed, but it is difficult to collect a large database for each task. We would therefore like to be able to use a relatively small database to adapt the model to a new task . This adaptation of statistical linguistic models is now under investigation.

2. Speech Recognition under Noisy Conditions

2.1 Overview

Recognition performance under noisy conditions is often impaired by variations in the amount of speech quality degradation rather than by the degradation itself. Problems are created, for example, by the variation of noise level associated with variations in the distance between the speaker and the microphone. If training can be performed under the same noise conditions as those under which the speech is to be recognized, performance is better than that attained when training occurred under noise-free conditions. Conversely, if training is performed under noisy conditions and the speech to be recognized is clean, performance is worse than when noisy speech is recognized.

If the range over which noise characteristics and speaking manner (such as speaking rate, loudness, and Lombard effect) are known ahead of time, training under the various expected conditions is useful [5]. This method is limited, however, because it is impossible to train under every condition. It would therefore be useful to have methods for automatically adapting to and normalizing the effects of adverse conditions.

Figure 1 shows the major methods that have been investigated [8], along with the basic sequence of acoustic processing in speech recognition. The following subsections explain the methods in each of these categories.

2.2 Special transducer arrangements

A head-mounted noise-canceling microphone (gradient microphone) is especially useful for dealing with the problems of low-frequency noise and room reverberation and for preventing the problems that occur when the distance between the speaker and the microphone varies [9]. Under very noisy conditions, like those in the cockpit of a fighter plane, this microphone is not effective enough. It has however been reported that the use of two-sensor input which combines an accelerometer (a throat microphone or a bone-contact microphone) output for low frequencies (< 1.5 kHz) and gradient microphone output for high frequencies (> 1.5 kHz) is effective [9]. The test reported, however, was limited to

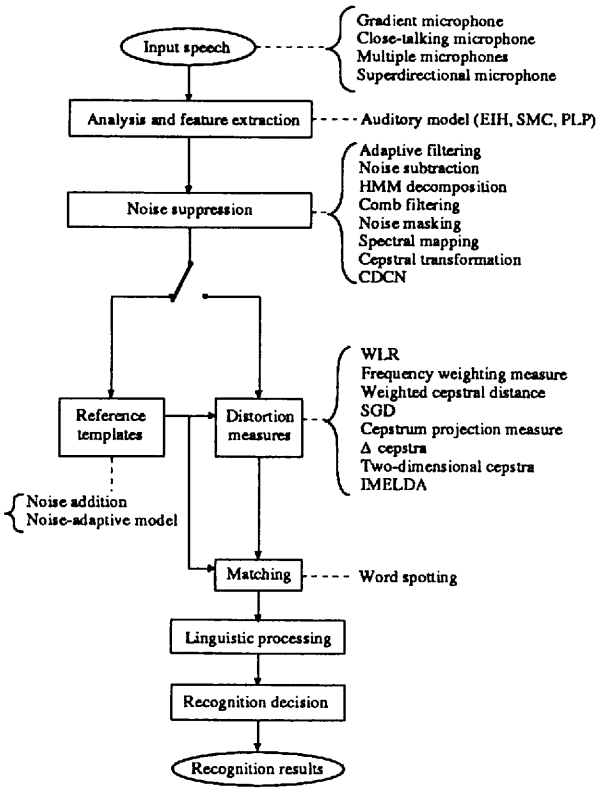

Figure 1. Major methods for coping with noise problems in speech recognition.

matched training and testing conditions, and thus cannot be extrapolated to unmatched situations.

Superdirectional microphones are useful when noise comes from a restricted direction. An adaptive microphone array for noise reduction (AMNOR) is a microphone array consisting of microphone units that are each followed by a digital filter. The characteristics of these digital filters are automatically adjusted to ensure that the microphone array is insensitive to sound coming from the direction of the noise source [10]. When the noise is diffused, however, as it is in airplanes and cars, the effectiveness of superdirectional microphones is relatively low [11].

One of the conventional methods for adaptive noise cancelation uses the adaptive filtering algorithm to process two input signals [11-12]. In this method (Fig. 2), the transfer characteristics from the noise source to the primary microphone are approximated by a transversal fil-

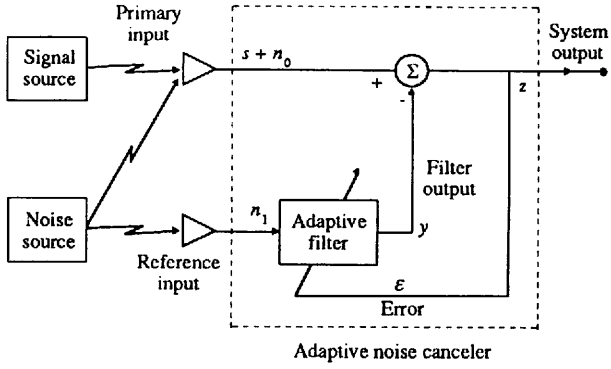

Figure 2. The concept of adaptive noise canceling (Reprinted from: B. Widrow, et al., Proc. IEEE, No. 12, 1975, p. 1693).

ter, and the noise at the primary microphone - estimated from the input of reference (auxiliary) microphone - is canceled. The noise that should be canceled must of course be properly estimated, and the larger the distance between the primary and reference microphones, the more difficult it is to estimate the noise. But if the distance between the microphones is small, it is difficult to prevent the speech signal from reaching the reference microphone. Although this method can reduce the fundamental wave component of car engine noise, a noise source that is easily captured, it is difficult to reduce diffused noise by using this method.

Adaptive noise cancelation was tested in a speech recognition experiment in a car. Speech period was determined from the power ratio of two microphone inputs, and noise masking and a special distance measure were applied. The distance measure consisted of inverse-variance weighted cepstra and delta-cepstra (see Subsection 2.6). For speech with an SNR of 13 to 24 dB, recognition was more accurate than when a conventional single-input method was used [13].

When speech recognizers are used for remote-control of acoustic equipment like stereo and television sets, the sound produced by the equipment becomes interfering noise. But because the noise is known in these situations, adaptive filtering is appropriate. It has been confirmed that an adaptive FIR filter controlled by a normalized LMS algorithm can reduce radio and television noise by 15 to 25 dB [14].

Yamada et al. [15] have investigated a method for removing reverberation by inverse filtering the band-

split reverberant speech obtained from multiple microphones. Because the construction of the inverse filter is based on MINT (multiple-input/output inverse) theorem [16] so as to minimize the difference between the inverse-filtered signal and reference signal for each frequency band, this method is called the sub-band MINT method (Fig. 3). When speech was uttered in a room with a reverberation time of 0.55 to 0.86 s and the sub-band MINT method was used with 512 sub-bands and two microphones, the original speech signal from 0.5 to 8.5 kHz was well recovered.

This method has been simplified to a multi-microphone sub-band selection method [17] and applied to word recognition. When a conventional single microphone was used, reverberation decreased recognition accuracy by roughly 10%, but when five to seven microphones were used and the inverse filtering was applied, the accuracy was decreased by only 2-3 percent.

2.3 Novel representations of speech (Auditory models)

The human hearing system is more robust than machine systems - more robust not only against the direct influence of additive noise but also against the speech variation (that is, the indirect influence of noise), even if the noise is very inconstant. A speech recognizer is therefore expected to become more robust when its front end uses models of human hearing. This can be done by imitating the physiological organ or by reproducing psychoacoustic characteristics.

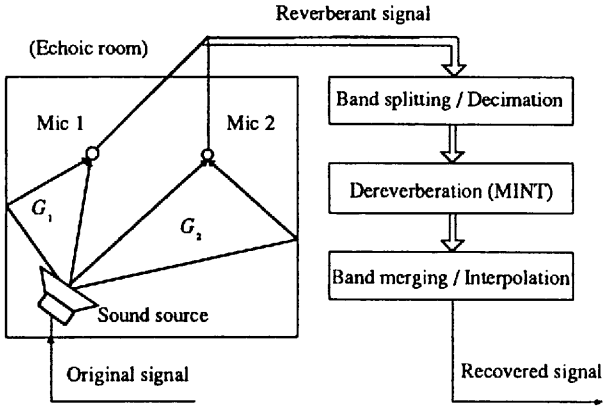

Figure 3. The sub-band MINT method (Reprinted from: Yamada et al., Proc. IEEE ICASSP, Toronto, 1991, p. 970).

The ensemble interval histogram (EIH) is a computational model based upon the temporal discharge patterns of auditory-nerve fibers [18]. It uses the ensemble histogram of interspike intervals, which histogram is generated by a simulated array of auditorynerve fibers, to produce a frequency-domain representation of the input signal (Fig. 4). The EIH model 1) uses 85 simulated cochlear filters splitting the frequency band between 200 and 3200 Hz, 2) detects the times when the output of each cochlear filter exceeds a threshold level, and 3) calculates histograms of level-crossing intervals. Although the histogram ensemble is similar to a frequency spectrum, it has a nonlinearity similar to that of human auditory processing characteristics and it has nonuniform frequency resolution. The recognition of noisy speech was greatly improved by applying this model to a DTW-based recognition system.

The short-time modified coherence (SMC) of speech increases the SNR of the speech signal by using the inherent coherence of adjacent segments in the signal [19]. Although the amount of improvement is data dependent, the SNR was increased by 10 to 12 dB when it was initially 0 to 20 dB. Using the SMC, speaker-dependent ten-digit recognition with noisy speech (whose SNR was 10 dB) provided a recognition accuracy of 98.3%, which was close to the 99.2% accuracy obtained with noise-free speech. When the traditional all-pole spectrum rep-

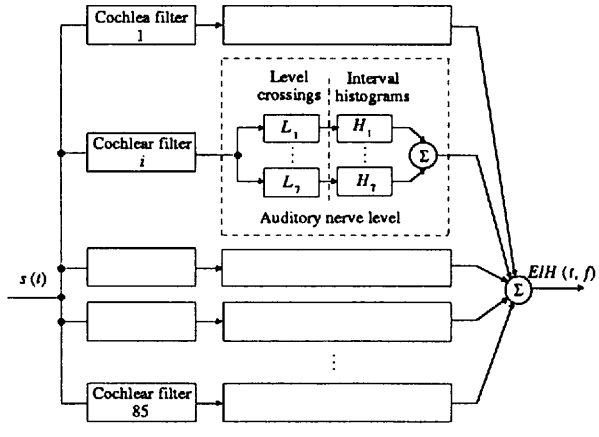

Figure 4. The ensemble interval histogram (EIH) computational model (Reprinted from: O. Ghitza, Journal of Phonetics, No. 16, 1988, p. 111).

resentation was used, the recognition accuracy with noisy speech was only 39.8%.

PLP (perceptual linear predictive) analysis is a technique that estimates the auditory spectrum by using three concepts derived from the psychophysics of hearing: 1) the critical-band spectral resolution, 2) the equal-loudness curve, and 3) the intensity-loudness power law. The auditory spectrum is then approximated by an autoregressive all-pole model [20]. As described in Subsection 2.6, the combination of PLP and weighted (liftered) cepstral distance measures is effective when recognizing noisy speech.

2.4 Signal enhancement preprocessing

Various research on noise suppression has been conducted [21]. Most noises are additive and can be classified according to whether they are correlated or uncorrelated to speech. Noises can also be classified as stationary or nonstationary. The most typical nonstationary noises are other voices. The difficulty of noise suppression depends on characteristics of the noise. A noise reference is often used for stationary noise, and one of the typical methods for suppressing stationary noise uncorrelated with speech is the Wiener filter method [22].

Ephraim et al. [23] have investigated a method for estimating the short-time noise level and the all-pole spectral model for the original speech from a given signal consisting of speech and uncorrelated additive noise. The estimation is performed by using iterative calculations that minimize the spectral difference between noise-added speech and a complex speech model. Let Y, σ/A, and N respectively be the power spectral density of noise-added speech, all-pole model spectral density, and noise power level. The values of σ/A and N are estimated by iterative calculations to minimize the Itakura-Saito distance measure defined by the following equation:

$$d(Y(\omega), \sigma/A(\omega) + N(\omega)) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\frac{Y(\omega)}{\sigma/A(\omega) + N(\omega)} - \log\frac{Y(\omega)}{\sigma/A(\omega) + N(\omega)} - 1\right\} d\omega \quad (1)$$

The estimated σ/A is used in speech recognition as a speech spectrum. An advantage of this method is that because it is unnecessary to know the noise level explicitly, this method can be used in the same way even when the noise level of recognition speech differs from that of training speech. The results of a speaker-dependent word recognition experiment using white-noise-added speech whose SNR was 10 dB showed that this method increased recognition accuracy from 37% to 72% when reference templates were made using clean speech.

Hidden Markov model (HMM) decomposition is a method for separating speech from a signal consisting of speech and additive noise [24-25]. Speech and noise are each modeled by a HMM, and the noisy signal is assumed to be modeled by a composite constructed by combining these two models. The Viterbi algorithm is used to decide which state sequence in the composite model is the most likely to have produced the observed signal. This corresponds to the process of searching a path in the three-dimensional space shown in Fig. 5. This method can be applied not only to stationary noise but also to time-varying noise, such as another speaker's voice. The effectiveness of this method was confirmed by experiments using speech signals to which noise or other speech had been added.

The cocktail party effect is the phenomenon that a person can extract one speaker's voice from noisy background. An engineering model of the cocktail party effect was constructed by using a comb filtering technique

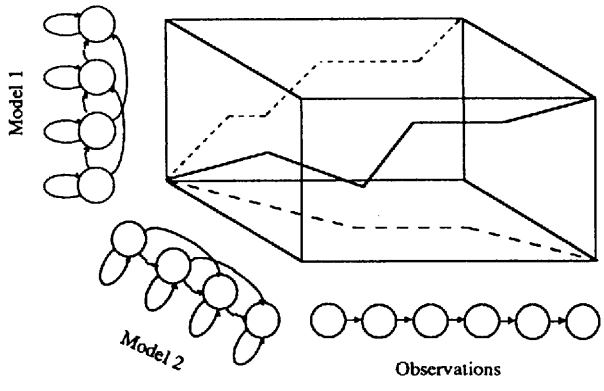

Figure 5. Decomposition of a 3-dimensional state sequence into two 2-dimensional projections in the Model 1 and Model 2 state spaces (Reprinted from: A. P. Varga et al., Proc. Eurospeech, 1991, p. 1176).

[26]. In an experiment using a signal consisting of two overlapped speech samples, the pitch frequency of one speaker's voice was extracted by using the temporal continuity of pitch frequency, and the speaker's voice was extracted by the comb filter controlled by the pitch frequency.

The comb filtering method was improved by adding the capability of estimating and subtracting the noise bias of each harmonic component [27]. A block diagram of this method is shown in Fig. 6. For speech with an SNR of 12 dB, this method increased the isolated word recognition rate from 25% to 80% in the case of white noise, and from 60% to 95% in the case of room noise.

Because it is not necessary in speech recognition to output a noise-reduced speech wave, methods for adding estimated noise to reference templates instead of reducing noise in the observed signal have also been investigated [28-29]. The noise-adding method is easier to use because it is free from the negative power problem that sometimes occurs in noise-reduction methods when noise level is over-estimated.

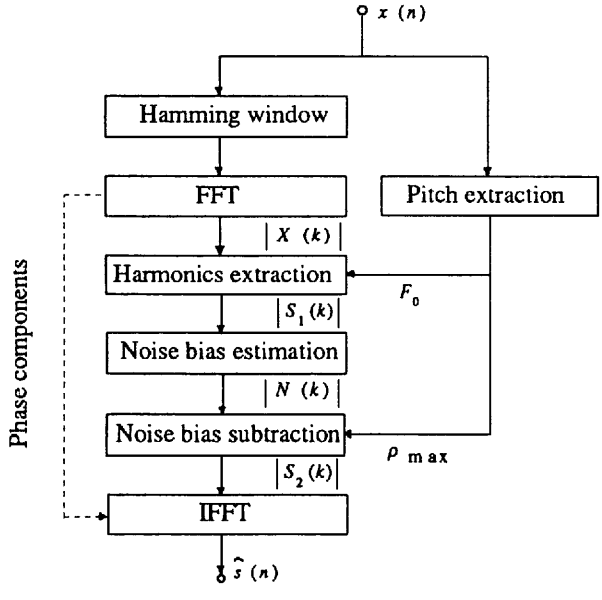

Figure 6. The comb filtering method for reducing noise (Reprinted from: H. Nagabuchi, Trans. IEICE Jpn., No. 5, 1988, p. 1101).

2.5 Noise masking and adaptive models

Klatt [30] tried combining filter-bank analysis and noise masking by a method in which only those frequency bands whose energy levels are higher than the masking levels derived from the noise levels are used in the distance calculation for each frame. The noise level of each band is estimated beforehand when there is no speech input. For each frame, the speech level of a frequency band in which the level of either the training utterance or the input speech is lower than the level of the masking noise is replaced by the masking level for both reference and input speech and therefore not used in the distance calculation. This method assumes that the noise level is stable and the noise level of input speech is not high. If the noise level is too high, a meaningful distance cannot be calculated because there are not enough frequency bands whose levels are higher than the masking levels. This method cannot be used directly in LPC-based speech recognition systems.

The noise masking method was also extended for application to stochastic-model-based recognition methods [31-32]. These methods combine noise models with HMMs, and experimental results have showed that one of these methods reduces recognition errors by roughly two-thirds [33-34].

Spectral mapping transforms noisy speech into clean speech by using correspondence rules between clean and noisy signals. The rules are obtained by using vector quantization techniques [35]. This method differs from most of the already mentioned methods, which use signal estimation theory to estimate the signal and the noise. It has been reported that applying the mapping method to a noisy signal with an SNR of 14 dB improved the SNR by roughly 10 dB.

A spectral mapping method based on hierarchical spectral clustering, which was originally proposed for speaker adaptation [36][7], has recently been applied to the spectral normalization of speech distorted by various additive noises, and its usefulness has been confirmed [37]. The use of neural nets for nonlinear mapping of cepstral parameters has also been investigated [38].

2.6 Robust distortion measures

Robust distortion measures selectively and automatically emphasize similarity (distortion) measures derived from those parts of the spectrum that are less corrupted

by noise. Because various methods described in the preceding subsections extract stable parts of noise-corrupted speech spectra or transform speech into stable representations, they can also be regarded - in a broad sense - as robust distortion measures.

Experimental evaluation of various distance measures in speaker-dependent word recognition based on LPC and filter-bank analysis have confirmed that the WLR measure weighted at the lower frequency bands is robust against white noise [39-41]. The WLR measure is defined by the following equation:

$$d_{WLR} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \{ \left(\frac{g(\omega)}{f(\omega)} - \log\frac{g(\omega)}{f(\omega)} - 1\right)\frac{f(\omega)}{\sigma^{(f)}} + \left(\frac{f(\omega)}{g(\omega)} - \log\frac{f(\omega)}{g(\omega)} - 1\right)\frac{g(\omega)}{\sigma^{(g)}} \} \, d\omega = 2\sum_{n=1}^{\infty} (r_n^{(f)} - r_n^{(g)})(c_n^{(f)} - c_n^{(g)}) \qquad (2)$$

$f(\omega)$: Reference spectral envelope

$g(\omega)$: Input spectral envelope

$\sigma^{(f)}$, $\sigma^{(g)}$: Power

$r_n^{(f)}$, $r_n^{(g)}$: Auto-correlation function

$c_n^{(f)}$, $c_n^{(g)}$: Cepstral coefficient

When the spectrum is represented by an all-pole model, this measure can be easily calculated by using autocorrelation functions and cepstral coefficients. The WLR measure increases robustness against noise by weighting spectral peak regions that are less affected by noises. A large-vocabulary isolated word recognition experiment evaluated the WLR measure by using speech that had an SNR of 18 dB and that was corrupted by white noise. With clean reference templates, the recognition rates were 85.0% when using the WLR and 65.3% when using cepstral distance measure [40].

An unsymmetrically weighted LR measure, in which noise-adaptive bandwidth expansion was combined with frequency weighting, has also been proposed [42-43].

Weighted (liftered) cepstral distance measures are generally represented by the following equation [44-45]:

$$d_{wcep} = \sum_{n=1}^{N} w^2(n)\,(c_n^{(f)} - c_n^{(g)})^2 \qquad (3)$$

Here $c_n^{(f)}$ and $c_n^{(g)}$ are the n-th cepstral coefficients for the two spectra that should be compared. It has been experimentally confirmed that since additive noises usually influence lower-order cepstral coefficients and the weighting factors, $w(n)$, deemphasize these coefficients, this measure improves the recognition accuracy for noisy speech [46-47].

The smoothed group delay (SGD) distance measure is a weighted cepstral distance measure in which the weighting factor is defined as $w(n)=n^s / \exp(-n^2/2t^2)$ ($s \geq 0$). Experiments with noisy or distorted speech have indicated that the most effective spectral peak emphasis is achieved by setting s between 1 and 2 and t at about 5. The recognition accuracy for speech with multiplicative noise and with an SNR of 10 dB was then 82%, whereas the accuracy without weighting was only 62% [48]. A weighted group delay (WGD) spectral distance measure that combines the advantages of the WLR and SGD measures has also been proposed and shown to be robust against noise [49].

Additive white noise reduces the norm (length) of the LPC cepstral vectors, and this norm shrinkage has been found to be a function of the noise level. The orientation (direction) of the cepstral vector is less susceptible to noise perturbation than is the vector norm. Although the shrinkage of the vector norm degrades recognition accuracy when conventional Euclidean distance calculations are used, the following cepstral projection measure is robust against differences between the noise levels of training and input speech. Its use results in recognition that is more accurate than when any other measure is used [50-52]:

$$d(C^{(f)}, C^{(g)}) = |C^{(g)}| \{1 - \frac{C^{*(f)} C^{(g)}}{|C^{(f)}||C^{(g)}|}\} \qquad (4)$$

Combination of this measure with the PLP analysis described in Subsection 2.3 is also effective for Lombard speech [53].

Delta-cepstra (or transitional cepstra), defined as regression coefficients for the time functions of cepstral coefficients over a rougly 50-ms period [54], are effective for recognizing noisy Lombard speech [55-56].

IMELDA (integrated mel-scale representation using linear discriminant analysis) is a distance measure based on the discriminant analysis technique, and its effectiveness for recognition of noisy speech has also been inves-

tigated [57-58].

Increased robustness against noise has also been achieved by using two-dimensional cepstral analysis for the frequency and time domains [59].

2.7 Stress compensation

Stress compensation is the process of compensating the spectral changes resulting from the vocal efforts made in order to adapt to noisy environments, and the spectral mapping technique described in Subsection 2.5 has been used for spectral-adaptation-based stress compensation [28]. In this method (Fig. 7), the spectra of reference templates made from stress-free speech uttered under the noise-free condition are mapped to the spectra of stressed speech uttered under the noisy condition.

To obtain the mapping rules, the same words are uttered under two conditions: while hearing noise through a headphone, and in a quiet environment. The quiet-environment speech is vector quantized, and a set of stressed speech spectra corresponding to each codebook element of stress-free speech is obtained by aligning the time axes of the two utterances. To obtain stress-compensated prototypes for use as the mapping targets of stress-free speech, each set of spectra is averaged in the autocorrelation domain. In addition, a noise spectrum is added to the prototypes in order to cope with the additive noise. For the isolated digit recognition task using speech utterances in a car running at 60 mph with the fan switched on, this technique reduced the recognition error rate from 29.9% to 9.6%.

Simple linear transformation of cepstral coefficients has also been used as a stress compensation technique

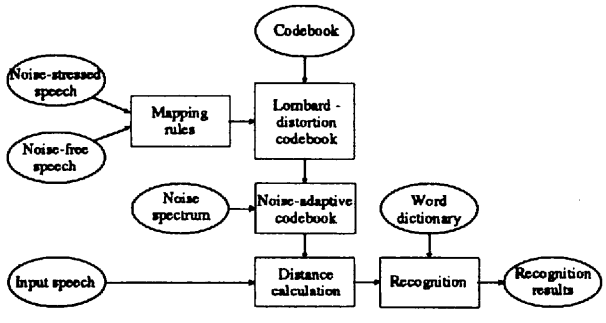

Figure 7. The noise-adaptive speech recognition system.

[60]. It was observed that the means and variances of cepstral vectors that define the observation probabilities of an HMM display some systematic modification in various speaking styles. Thus, a compensated word model could be constructed by shifting the means and scaling the variances in the original HMM according to the observed modifications. In isolated-word recognition experiments including six speaking styles, this method reduced the substitution error rate from 25.9% to 16.4%.

3. Coping with the Speech Variation Resulting from Different Kinds of Microphones

The difference between the characteristics of microphones used for training and recognition sometimes greatly degrades recognition. For example, in the speaker-independent recognition of continuous speech with a perplexity of 65, the recognition accuracy was 85% when test utterances were obtained through the same close-talking microphone used for recording training utterances, whereas it decreased to less than 19% when the test utterances were obtained through a desk-mounted microphone [61].

The effectiveness of the CDCN (codeword-dependent cepstral normalization) method proposed to cope with this problem has been confirmed by various experiments [62-64]. In this method, noise level and spectral inclination are estimated directly from input speech, without using training utterances, in the cepstral domain. When the distortion characteristics for the speech signal $x(t)$ are represented by the linear filter $h(t)$ and the uncorrelated noise added to the output of this filter is represented by $n(t)$, power spectral density is represented as

$$Y_z(w) = Y_x(w)\,|H(w)|^2 + Y_n(w) \qquad (5)$$

Although it is difficult to directly and separately extract the effects of distortion and noise from $Y_z(w)$ in the cepstral domain, by introducing several approximations the noise and spectral inclination there can be estimated under the maximum likelihood criterion. The estimation is executed by iterative calculations similar to those in the EM algorithm. Using the CDCN method is more effective than independently applying spectral subtraction and spectral normalization methods.

In the speaker independent recognition of continuous speech with a perplexity of 65, the recognition rate when using the different microphone was almost the same as that achieved when using the same microphone [64]. Another advantage of the CDCN method is that it can adapt dynamically to the variation of acoustical environment, since it does not need long-term a priori statistical information about noise.

4. Coping with Spontaneous-Speech-Specific Problems

As described in Section 1, the recognition of spontaneous speech poses a lot of difficult problems that do not occur in the recognition of speech that is read. A word spotting technique is frequently used to determine the speech period in spontaneous speech - that is, for separating speech from stationary and nonstationary noises in real environment. The nonstationary noise includes such speech and speech-like sounds as coughing, restarting, and the repetition of utterances. The word spotting method extracts these words from continuous speech by using acoustical and temporal characteristics of target words in the top-down manner [65-66].

Systems for recognizing spontaneous speech should tolerate "barge-in" - that is, talk-over-prompt speech input. Allowing people to talk over prompts, thus short-cutting instructions, is essential to constructing a comfortable interactive system. This means that we have to be able to deal with speech-on-speech issues. If the prompting speech is known, the easiest solution is to completely isolate the input and output channels and use echo cancelation technology. A more challenging method, which was described in Subsection 2.4, is to model speech in the presence of speech.

In spontaneous speech, correct sentences are not always spoken, and obstacle words and sounds that are unnecessary for understanding the utterances are frequently added. Additionally, the system is requested to respond to the utterance as quickly as possible. To solve these problems, it is necessary to establish a method for detecting the time at which sufficient information has been acquired instead of detecting the end of input speech.

Style shifting is also an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers are reading lists of words rather than trying to accomplish a real task. Users actually trying to accomplish task, however, use a different linguistic style. This task-oriented style is hard to simulate precisely, but ATIS data collection methods tried in the DARPA community seem to come closest [67].

5. Speaker Adaptation

Approaches to building speaker-independent recognition systems include [7]:

(1) Signal processing to extract invariant features
(2) Information theoretic approaches (such as those using discriminant methods)
(3) Pattern recognition (including multiple-template methods)
(4) Artificial intelligence (knowledge engineering)
(5) Statistical approaches (for example, using HMMs)
(6) Neural networks

The performance of speaker-independent speech recognition systems has recently been greatly improved by training them with a large speech database spoken by many speakers and by incorporating statistical models of speech variations. It is still impossible, however, to accurately recognize the utterances of every speaker. Experiments have shown that people can adapt to a new speaker's voice after hearing just a few syllables [68]. Recent research has therefore explored the possibility of giving recognition systems the ability to automatically adapt to individual speakers [6].

Speaker-adaptation methods are generally classified either as supervised (text-dependent) methods, in which training words or sentences are known, or as unsupervised (text-independent) methods, in which arbitrary utterances can be used. The major speaker-adaptation methods which have recently been investigated are the following:

(1) Speaker cluster selection methods
(2) Interpolated re-estimation algorithm
(3) Codebook adaptation/normalization algorithm
(4) Reference-templates generation methods

It is essential to improve the performance of speaker-independent recognition systems and to develop a method that automatically adapts these systems to a new speaker. The adaptation method should be able to increase performance to match that of the speaker-dependent systems - that is, systems using a large database of

training utterances uttered by the speaker whose speech is being recognized [69].

6. Adaptation to a New Task

Speech recognition systems basically consist of acoustic and linguistic models, and both models change their characteristics according to the recognition tasks. For acoustic models, task-independent phone models have recently been explored [70]. But because automatic training using utterances without phone labels is now possible [71], it is not difficult to use large speech databases to make task-dependent acoustic models that can be changed every time a task changes.

The adaptation of linguistic models, in contrast, is still a very important issue. This is because a linguistic database needs to be much bigger than an acoustic database, and collecting a large linguistic database for a new task is difficult and costly. Various research has therefore explored the adaptation of linguistic models [72-74].

7. Conclusion

This paper has briefly reviewed the major methods for improving the performance of speech recognition systems used under field conditions - when speech signals are subject to many different kinds of variations. Making speech recognition systems that are robust when used under adverse field conditions requires us to answer a fundamental question: what the essential feature parameters and methods should be. That is, methods that are not robust in actual use cannot be considered to be essential methods.

Although many methods for robust speech recognition under adverse conditions have already been proposed, these methods have not yet been fairly compared. It will therefore be essential to collect databases under field conditions (actual or simulated) and to use these databases to compare the effectiveness of the proposed methods. It is also necessary to make clear the appropriate application areas for each major method, to set forth guidelines for combining these methods, and to improve the most promising methods.

To correctly evaluate speech recognition technologies and to achieve steady technical progress, it is important to comprehensively evaluate a technology under actual field conditions instead of under a single controlled laboratory condition. It is also important to evaluate and improve the technology from the viewpoint of the human interface, for which evaluation under actual and adverse conditions is essential.

References

1. D. B. Pisoni, R. H. Bernacki, H. C. Nusbaum and M. Yuchtman, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tampa S41.10 (1985) 1581.
2. I. Lecomte, M. Lever, J. Boudy and A. Tassy, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow S10a.12 (1989) 512.
3. J. C. Junqua and Y. Anglade, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque S15b.9 (1990) 841.
4. P. K. Rajasekaran, G. R. Doddington and J. W. Picone, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 14.10 (1986) 733.
5. R. P. Lippmann, E. A. Martin and D. B. Paul, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 17.4 (1987) 705.
6. S. Furui, Speech Communication Nos. 5-6 (1991) 505.
7. S. Furui, in Advances in Speech Signal Processing, ed. by S. Furui and M. M. Sondhi (Mercel Dekker, New York) (1992) 597.
8. B. H. Juang, Proc. Int. Conf. Spoken Language Processing, Kobe 25.1 (1990) 1113.
9. V. Viswanathan, C. Henry, R. Schwartz and S. Roucos, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 3.2 (1986) 85.
10. Y. Kaneda and J. Ohga, IEEE Trans. Acoustics, Speech, Signal Processing No. 6 (1986) 1391.
11. G. A. Powell, P. Darlington and P. D. Wheeler, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 6.2 (1987) 173.
12. B. Widrow, J. R. Glover, Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, Jr. and R. C. Goodlin, Proc. IEEE No. 12 (1975) 1692.
13. Y. Nakadai and N. Sugamura, Proc. Int. Conf. Spoken Language Processing, Kobe 25.8 (1990) 1141.
14. T. Usagawa, Y. Morita and M. Ebata, Proc. Korea-Japan Joint Symposium on Acoustics S1-3-J3 (1991) 222.

15. H. Yamada, H. Wang and F. Itakura, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.23 (1991) 969.
16. M. Miyoshi and Y. Kaneda, IEEE Trans. Acoustics, Speech, Signal Processing No. 2 (1988) 145.
17. H. Wang and F. Itakura, IEICE Technical Report SP91-62 (1991) 15.
18. O. Ghitza, Journal of Phonetics No. 16 (1988) 109.
19. D. Mansour and B. H. Juang, IEEE Trans. Acoust., Speech, Signal Processing, No. 6 (1989) 795.
20. H. Hermansky, J. Acoust., Soc., Am. No. 4 (1990) 1738.
21. S. F. Boll, in Advances in Speech Signal Processing, ed. by S. Furui and M. M. Sondhi (Mercel Dekker, New York) (1992) 309.
22. A. D. Berstein and I. D. Shallom, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.9 (1991) 913.
23. Y. Ephraim, J. G. Wilpon and L. R. Rabiner, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 31.1 (1987) 1324.
24. A. P. Varga and R. K. Moore, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque S15b.10 (1990) 845.
25. A. P. Varga and R. K. Moore, Proc. Eurospeech 91 (1991) 1175.
26. T. W. Parsons, J. Acoust., Soc., Am. No. 4 (1976) 911.
27. H. Nagabuchi, Trans. IEICE Jpn. No. 5 (1988) 1100.
28. D. B. Roe, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 27.5 (1987) 1139.
29. Y. Takebayashi, H.Tsuboi and H. Kanazawa, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.7 (1991) 905.
30. D. H. Klatt, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Philadelphia (1976) 573.
31. J. N. Holmes and N. C. Sedgwick, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 4.12 (1986) 741.
32. D. V. Compernolle, Computer Speech and Language No. 3 (1989) 151.
33. A. Nadas, D. Nahamoo and M. A. Picheny, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, New York S11.8 (1988) 517.
34. V. L. Beattie and S. J. Young, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.10 (1991) 917.
35. B. H. Juang and L. R. Rabiner, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 6.6 (1987) 2368.
36. Y. Shiraki and M. Hoda, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque S12.5 (1990) 657.
37. H. M. Cung and Y. Normandin, Proc. ESCA Workshop on speech processing in adverse conditions (1992).
38. H. B. D. Sorensen, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.14 (1991) 933.
39. K. Shikano and F. Itakura, in Advances in Speech Signal Processing, ed. by S. Furui and M. M. Sondhi (Mercel Dekker, New York) (1992) 419.
40. M. Sugiyama and K. Shikano, Trans. IEICE Jpn. No. 4 (1983) 385.
41. H. Matsumoto and H. Imai, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 14.19 (1986) 769.
42. F. K. Soong and M. M. Sondhi, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 15.1 (1987) 625.
43. F. K. Soong and M. M. Sondhi, IEEE Trans. Acoust., Speech, Signal Processing No. 1 (1988) 41.
44. S. Furui, IEEE Trans. Acoust., Speech, Signal Processing No. 2 (1981) 254.
45. Y. Tohkura, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 14.17 (1986) 761.
46. B. H. Juang, L. R. Rabiner and J. G. Wilpon, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 14.18 (1986) 765.
47. B. A. Hanson and H. Wakita, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo 14.16 (1986) 757.
48. F. Itakura and T. Umezaki, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 29.1 (1987) 1257.
49. H. Matsumoto and H. Mitsui, Trans. IEICE Jpn. No. 8 (1991) 1257.
50. D. Mansour and B. H. Juang, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, New York S1.5 (1988) 36.
51. D. Mansour and B. H. Juang, IEEE Trans. Acoust., Speech, Signal Processing No. 11 (1989) 1659.

52. B. A. Carlson and M. A. Clements, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.11 (1991) 921.
53. J. C. Junqua and H. Wakita, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow S10a.3 (1989) 476.
54. S. Furui, IEEE Trans. Acoust., Speech, Signal Processing No. 1 (1986) 52.
55. B. A. Hanson and T. H. Applebaum, Proc. Int. Conf. Spoken Language Processing, Kobe 25.2 (1990) 1117.
56. T. H. Applebaum and B. A. Hanson, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.27 (1991) 985.
57. M. J. Hunt and C. Lefebvre, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow S6.3 (1989) 262.
58. M. J. Hunt, S. M. Richardson, D. C. Bateman and A. Piau, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.1 (1991) 881.
59. Y. Ariki, S. Mizuta and T. Sakai, Trans. IEICE Jpn. No. 5 (1988) 790.
60. Y. Chen, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas 17.7 (1987) 717.
61. A. Acero and R. M. Stern, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque S15b.11 (1990) 849.
62. A. Acero and R. M. Stern, Proc. Int. Conf. Spoken Language Processing, Kobe 25.3 (1990) 1121.
63. A. Acero and R. M. Stern, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.4 (1991) 893.
64. R. M. Stern, F. H. Liu, Y. Ohshima, T. Sullivan and A. Acero, Fifth DARPA Workshop on Speech & Natural Language Session 8b (1992).
65. J. G. Wilpon, L. R. Rabiner and T. Martin, AT&T Bell Laboratories Technical J. No. 3 (1984) 479.
66. A. Imamura and Y. Suzuki, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S13.5 (1991) 537.
67. L. Hirschman, Fifth DARPA Workshop on Speech & Natural Language Session 1 (1992).
68. K. Kato and S. Furui, IEICE Technical Report H85-5 (1985).
69. J. L. Gauvain and C. H. Lee, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Francisco S56.3 (1992) I-481.
70. H. W. Hon and K. F. Lee, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Toronto S14.3 (1991) 889.
71. K. F. Lee, Automatic Speech Recognition - The Development of SPHINX System, Kluwer Academic Publishers, Boston (1989).
72. R. Kuhn and R. De Mori, IEEE Trans. Pattern Analysis and Machine Intelligence No. 6 (1990) 570.
73. S. Matsunaga, T. Yamada and K. Shikano, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Francisco S25.3 (1992) I-165.
74. S. D. Pietra, V. D. Pietra, R. L. Mercer and S. Roukos, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Francisco S25.4 (1992) I-633.