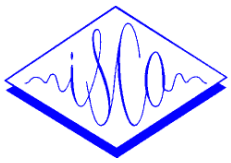

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Recent Developments in Speech Recognition under Adverse Conditions

B. H. Juang

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

In this paper, we review several promising methods that were proposed in the past few years to deal with the problem of speech recognition in adverse conditions. We discuss these methods in six categories: signal enhancement preprocessing, special transducer arrangements, noise masking, stress compensation, robust distortion measures, and novel speech representations. We explain each categorical approach and provide a digest of the performance improvements each method is able to achieve. This type of information is helpful in making a technical decision for the actual recognizer design to be deployed in adverse environments.

1.0 Introduction

The problem of automatic speech recognition in an adverse environment has attracted the attention of many researchers. The main reason is that the performance of existing speech recognition systems, whose designs are based on low noise or low distortion input, degrades rapidly in the presence of noise and distortion.

In their study of noise effects alone, Dautrich et. al. [1] demonstrated that an isolated word recognizer trained in clean (virtually noise free) conditions and capable of achieving a recognition accuracy of 95% had an order of magnitude increase in error rate when tested with noisy utterances at an SNR (signal-to-noise ratio) of 18 dB. The problem requires more attentive solutions than a simple noisy training because training reference patterns under the exact matched noisy test condition is seldom affordable and the use of reference patterns trained on noisy input leads to unacceptable results when the test condition is actually clean.

Recently, several methods and algorithms have been proposed to specifically deal with the adverse environment in which the speech recognizer is to be deployed. The goal is to have an automatic recognizer with robust performance approaching that of matched conditions (as if the recognizer were trained under the test condition). Here, we review these methods and point out their algorithmic characteristics.

2.0 Adverse Conditions in Speech Recognition

2.1 Noise

Acoustic ambient noise is usually considered additive, i.e. the recorded signal is a sum of the intended speech signal and the ambient noise. Sources of acoustic ambient noise are abundant. In the office environment, the office machinery such as a typewriter or printer, personal computers or workstations which are usually equipped with moving components like disks and fans, telephone ringing and background conversation of other people, etc. often emit enough acoustic noise to cause performance degradation of a speech recognizer. The sound pressure level (SPL) in a normal personal office is around 45 ~ 50 dBA (noise criterion 40 ~ 45). Fig. 1 shows a typical spectrum, fitted with a 16th order all-pole smoothed spectrum, of ambient noise recorded in a personal office with a SUN 3/110 on and operating. In a business office where secretarial duties are performed, the SPL could be 15-20 dB higher than the above figure. Inside an automobile, the acoustic noise level due to engine, cooling fan, wind, tire and road is usually even higher, particularly when the automobile is in motion. It was found [2,3] that the SNR of speech signals recorded in a car with a microphone mounted on the dashboard in front of the speaker/driver could drop below −5 dB when the car was cruising at a speed of 90 km/h. In the cockpit of a modern jet fighter aircraft, SPL's of 90 dB or more across the speech frequency band have been reported [4]. The spectrum of acoustic ambient noise, as seen in Fig. 1, is usually not flat.

Other types of noise such as electrical noise and quantization noise, which are present in any modern automatic speech recognition system, are in general at a level below the threshold of concern. Nevertheless, noise due to transmission line and switching equipment in a telephone network can sometimes be substantial [5].

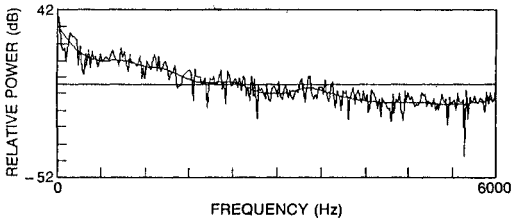

Fig. 1. Noise spectrum in a personal office with a SUN 3/110.

2.2 Distortion

Aside from the additive contamination due to noise like signals, the uttered signal inevitably undergoes a series of spectral distortions before being processed for speech recognition. The room in which the speech recognizer is deployed almost certainly has a varying degree of reverberation that can alter the signal spectrum. The microphone transducer, depending on its type and mounting position, also can significantly distort the speech spectrum. In case that the transducer used in testing is different from the one used during training, the mismatch in spectral distortion becomes one of the major problems. For example, it was reported [6] that a large vocabulary speech recognition system with a baseline performance of 85% accuracy in a matched transducer condition (Sennheiser HMD224 close-talking microphone for both training and testing) could only achieve less than 19% accuracy when a different microphone (Crown PZM6fs desk-top mounted) was used during testing.

When the automatic recognizer is deployed in a telephone network, the telephone channel can cause further distortion of the signal spectrum. A list of statistics of attenuation distortion relative to 1004 Hz for short end-office connections (0-180 airline miles) as reported in [5] clearly indicate this concern. For example, at 204 Hz, the mean attenuation is 5.1 dB, but at the 10% and 90% quantile, the attenuation is 2.5 dB and 10 dB respectively. And at 3204 Hz, the mean attenuation is 4.1 dB, with 0.5 dB and 11.5 dB attenuation at the 10% and 90% quantile, respectively. This wide range of variation in attenuation will cause spectral mismatch distortion unless the telephone channel is measured/learned before every recognition trial.

2.3 Articulation Effects

Many factors affect the talker in his/her speaking manner. Even the psychological awareness of communicating with a speech recognizer could make the talker produce a noticeable difference in his/her formants and rhythmic stability [2]. Characteristic changes in articulation due to environmental influence, known as the Lombard effect, can be dramatic [2,3,7-9]. When a talker speaks under a masking noise of 90 dB SPL, Pisoni et. al. reported [7] that the first formant of a vowel often increases while the second formant decreases. In a separate study [8], Junqua et. al. estimated the increase in the first formant frequency due to an 85 dB SPL masking noise to be between 42 and 113 Hz. Furthermore, a significant change in spectral tilt was also observed [8]. These characteristic changes dramatically affect the performance of an automatic speech recognizer. A speaker-dependent isolated word recognizer that has an accuracy of better than 92% when the training and the testing are conducted under the same (virtually free from noise and Lombard effect) condition can only maintain an accuracy of 61% if the test utterances contain the Lombard effect, even though they are free from masking noise [10].

A major difficulty with the varying articulation effects is to quantify them. With acoustic noise or channel distortion which usually does not vary as rapidly as the speech itself in terms of spectral characteristics, we can to some extent measure or model them to help the recognizer design. Articulation effects, however, are a byproduct of the speech production process and are found to be context dependent [8].

3.0 How to Deal with Adverse Conditions

If the characteristics of the corrupting noise are to some extent known, a speech recognizer that uses reference patterns trained with noisy speech in general performs more robustly than one that uses clean reference patterns. This can be extended to other cases where the talker, due to his/her psychological reaction to environmental stress, produces speech in unusual talking styles (slow, fast, soft, loud, angry and Lombard). The idea is simply to train the recognizer with a multi-style training procedure in which speech signals of different talking styles are used as the training data. Lippmann et. al. [10] has demonstrated this possibility.

We have argued, however, that training material reflecting actual deployment conditions usually is not readily available. Straightforward solutions like noisy or multi-style training may not satisfy the problem of robust speech recognition in adverse conditions. In the following, we discuss a number of methods and algorithms that have been proposed to combat the often unknown and severe environment a speech recognizer faces in practice.

3.1 Signal Enhancement Preprocessing

When the adverse condition is due to additive noise alone, one can employ speech enhancement methods to suppress the noise before applying the recognition algorithm. One of the most widely studied signal enhancement methods is adaptive noise cancellation using two signal sources [11]. This technique, however, requires that the noise component in the corrupted signal and the reference noise have high coherence. Inside a passenger car, it was found [12] that if the two microphones are located at a distance greater than 50 cm, the only coherent noise component is that of the engine. In order to cancel 90% of the noise energy, the two microphones cannot be more than 5 cm apart, which makes it almost impossible to prevent speech from entering the noise reference. Therefore, one can only expect cancellation of noise which is related to the fundamental frequency of the engine revolution if the two microphones are at a reasonable distance.

Other speech enhancement techniques that do not rely on the existence of a simultaneous, separate noise reference have also been considered [14-17 and references therein]. The least squares estimator of short time independent spectral components of [15] differs from a traditional estimator [18] in that the conditional mean of the spectral component is obtained from the sample average estimator of clean speech rather than from an assumed parametric distribution. The method uses a clean speech data base and a noisy version of it by artificially adding noise to the clean set to construct a function that maps a noisy spectral component (at each frequency) to a noise-suppressed value. Under the condition that the signal and the noise levels are fixed and known (SNR = 10 dB), the technique was reported to be able to reduce the recognition errors due to noise by three quarters (from ~40% to ~10%) for a speaker dependent digit recognition task.

The method of [17] is interesting in that the short time noise level as well as the short time all-pole spectral model of the clean speech are iteratively estimated to minimize the Itakura-Saito distortion between the noisy spectrum and a composite model spectrum. That is, if $|Z|^2$, $\sigma^2/|A|^2$, and λ represent the noisy power spectral density, the all-pole model spectral density and the noise power level respectively, the algorithm iteratively finds σ/A and λ such that

$$d(|Z|^2, \sigma^2/|A|^2+\lambda) =$$

$$\int_0^{2\pi} \left[\frac{|Z(\omega)|^2}{\sigma^2/|A(\omega)|^2 + \lambda} - \ln \frac{|Z(\omega)|^2}{\sigma^2/|A(\omega)|^2 + \lambda} - 1 \right] \frac{d\omega}{2\pi}$$

is minimized. The resultant σ/A is then used in the recognizer as the spectral measurement of speech. The method can be applied to test utterances as well as training utterances without explicit knowledge of the noise level. In a speaker dependent isolated word recognition experiment, the technique significantly improved the recognition accuracy, from 42% when unprocessed clean reference templates were used for 10 dB SNR test tokens to almost 70% when both the reference and the noisy test tokens were processed to produce the enhanced spectral measurement.

3.2 Special Transducer Arrangements

If the talker position is fixed, a noise cancelling microphone can be effective in suppressing low frequency noises in an automobile or aircraft cockpit environment. In a passenger car using a pressure gradient noise cancelling microphone (CONFIDENCER by Roanwell), Dal Degan and Prati [12] confirmed that the signal picked up is essentially noise free if the microphone is kept very close to the talker's mouth and parallel to the wavefront. But with a mere 10 cm shift and 30-degree rotation, the speech power drops by 15 dB, resulting in a performance degradation.

Other gradient microphones were also found by Viswanathan et. al. [20] to be effective in a moderately noisy environment (95 dB SPL broadband acoustic noise) if the sensor location is optimized and fixed. When the noise is severe as in a fighter aircraft cockpit at 105 dB SPL, Viswanathan [20] reported that the noise cancelling microphone alone did not lead to satisfactory recognition results and suggested the use of two-sensor input which combines an accelerometer output for low frequencies (up to ~1.5K Hz) and gradient microphone output for high frequencies (above ~1.5K Hz). Another possibility also suggested in [20] is to use the two sensor outputs in parallel to facilitate composite feature extraction. Although performance improvements were shown with various multi-sensor arrangements as compared to each single constituent transducer, the test was limited to matched training and

testing conditions and thus cannot be extrapolated to the mismatched situation where the characteristics of the ambient noise may be varying due to changing flying speed and altitude.

3.3 Noise Masking and Adaptive Models

In the presence of broadband noise, certain regions of the speech spectrum that are of lower level will be more affected by the noise. This makes the calculation of spectral distortion difficult since those more corrupted regions represent less reliable spectral measurements. Recognizing this difficulty, Klatt [21] advocated use of noise masking in conjunction with a filter-bank analyzer. The key idea of Klatt is to first choose for each channel the noise mask level as the higher of that in the reference signal and that in the testing signal, and then replace the channel output by the mask value if it is below the corresponding mask level. This helps prevent spurious distortion accumulation.

Klatt's masking scheme, however, has practical limitations, particularly when the two patterns being compared have very different noise levels. When the test token is contaminated by a high level of noise, all the reference patterns that are of lower level than the noise would result in equally small distances making the comparison meaningless. Subsequent improvements to Klatt's technique were proposed by Bridle et. al. [22] and Holmes and Sedgwick [23]. These revisions incorporate the following to overcome the above limitation: 1) maintain a running estimate of the noise spectrum, both during training and testing; 2) separately mark the channel spectral values of the training and the testing tokens as speech or noise according to respective noise estimates; 3) devise individual distance calculation rules for different marking situations; and 4) in case of multiple reference patterns, use the maximum of the noise estimates for the particular channel during training.

An attempt to incorporate noise masking in a probabilistic modeling framework was also made by Holmes and Sedgwick [23], followed by Van Camperolle [24] and Nadas et. al. [25]. Let random variables X, Y, and Z be the clean speech spectrum, the noise model spectrum, and the observed noisy spectrum respectively. (Note that $Z=z$ is the only observed value during test.) The masking procedure is effectively modeling Z as $Z = \max(X,Y)$. The pdf of Z, say h, can be conveniently expressed in terms of the distributions of X and Y and can be used to approximate the true noisy pdf. Thus, similar to the design of a vector quantizer (VQ) codebook, a set of $h(z)$'s can be designed as the noise-compensated prototypes of the speech given the noise estimate. These prototypes are then used to facilitate labeling of each input noisy spectral vector for further HMM modeling. In [25], the noise was produced in an office environment and the "clean" and "noisy" conditions were associated with SNR of 41 dB and 31 dB respectively. It was demonstrated that noise compensation was able to reduce the errors by two thirds from ~32% using clean prototypes for noisy test tokens to ~10% with the noise-adapted prototypes. The technique of noise compensation was also employed by Roe [3] to adapt the spectral prototypes to the noise condition in the autocorrelation domain.

3.4 Stress Compensation

Stress compensation is to provide offset for the spectral distortion caused by unusual speaking effort due to the talker's reaction to ambient conditions. The recognizer of Roe [3] is a traditional template-based system that incorporates both VQ and dynamic time warping (DTW). In the system, the spectral pattern is first vector quantized and replaced by the closest spectral prototype in the VQ codebook for subsequent distortion calculation. The reference template thus consists of a sequence of VQ spectral prototypes. To adapt the "stress-free" clean spectral prototypes to the stressed speech, each (known) stressed speech utterance is time-aligned, without VQ, to the correct reference template. The autocorrelation vectors of the stressed speech frames are then grouped according to the corresponding VQ indices in the reference templates and averaged to yield the stress-compensated prototypes. In a speaker-trained isolated digit recognition trial, the stress compensation scheme reduced the number of recognition errors by two thirds, from 29.9% to 9.6%, when the extra speaking effort was caused by the noise in a car cruising at 60 mph with fan on.

Another stress compensation technique operating in the cepstral domain was proposed by Chen [26]. The basic presupposition of the technique is that spectral distortion induced by unusual speaking efforts can be compensated by simple linear transformation of the cepstrum. Although this presupposition may be unduely strong, 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 a speaker dependent isolated word recognition trial involving six different speaking styles [26], the compensated models were able to reduce the error rate (substitution only) from 25.9%, when only the normal models were employed, to 16.4%.

3.5 Robust Distortion Measures

Spectral weighting for improving speech recognition accuracy has long been considered (e.g. [27]) a viable approach. The work of Matsumoto and Imai [28] represents an early effort to investigate the sensitivity of various weighted distortion measures to noise contamination in recognition tasks. The weighted measures considered in [27] and [28] are based on the log spectral difference. Let $|X(\omega)|^2$ and $|Z(\omega)|^2$ be two power spectra subject to comparison. The log spectral difference is $V = \log|X|^2 - \log|Z|^2$. Many weighted distortion measures can be accordingly defined, e.g.

1. Weighted Likelihood Ratio (LR) Distortion:
$$\int_0^{2\pi} \left[(V+e^V-1)\frac{|X|^2}{P_X} + (-V+e^{-V}-1)\frac{|Z|^2}{P_Z} \right] \frac{d\omega}{2\pi}$$
2. Asymmetrically Weighted LR Distortion:
$$\int_0^{2\pi} (V+e^V-1)\frac{|X|^2}{P_X}\frac{d\omega}{2\pi}$$

where P_X and P_Z are the average power of X and Z respectively. When X and/or Z are all-pole models, these distortion measures can be easily calculated in terms of the autocorrelation and the cepstral coefficients. The weighting is obviously accomplished by boosting the spectral peak regions which have more power concentration to resist noise corruption. It was reported that at 18 dB SNR (white noise) using only clean reference templates, weighted measures greatly improved the recognition performance (from 60% to 90%) in a speaker dependent isolated word (28 Japanese city names) recognition trial. Soong and Sondhi [29] also proposed a weighted measure, similar to the above asymmetrically weighted LR, with a noise-adaptive bandwidth expansion factor in the weighting spectrum.

Weighted (liftered) cepstral distance measures have the form $d = \sum w^2(n)(c(n) - c'(n))^2$ where $c(n)$ and $c'(n)$ are the cepstral coefficients of the two spectra being compared. When proper liftering functions $w(n)$ are used, weighted cepstral

measures have been found to be advantageous for speech recognition in clean as well as noisy conditions [30-32]. The work of Itakura and Umezaki [32] is a comprehensive study of weighted cepstral measures in noisy and distorted conditions. The cepstral lifter $w(n)$ considered in [32] has a general form of $w(n) = n^s \exp(-n^2/2\tau^2)$ $s \geq 0$. Several adverse test conditions were considered for a set of confusing Japanese city name pairs. The cepstral liftering led to various degrees of performance improvements; for example, at 10 dB SNR (multiplicative noise), the recognition accuracy was enhanced from ~62% without cepstral liftering to ~82% with cepstral liftering for $s=1.5$ or 2. Overall performance for various conditions was found to be best with parameters $s=1 \sim 2$ and $\tau=5$. It is interesting to note that with these parameters the smoothed group delay weighting function is very close to the raised sine lifter of [30].

The pursuit of a robust distortion measure can be more effective if we have an analytical understanding of the effects of noise on spectral parameters. Mansour and Juang [33] reported that analytical and experimental evidences indicate that additive white noise causes the cepstral vector norm (length of the cepstral vector, without the zeroth term) to shrink but leaves the cepstral vector orientation more or less intact. The vector norm shrinkage is detrimental in the traditional Euclidean distance calculation. Also, since the norm shrinkage was found to be a function of the noise level, the vector norm itself can be used to facilitate non-uniform weighting for each speech frame in the accumulative distance. The result suggests the use of a projection operation to formulate several distortion measures to cope with the mismatched noisy condition. In [33], the following cepstral projection measure was found to be a good choice overall:

$$d(\mathbb{C}_r, \mathbb{C}_t) = |\mathbb{C}_t| \left[1 - \frac{\mathbb{C}_r^* \mathbb{C}_t}{|\mathbb{C}_r| \, |\mathbb{C}_t|} \right]$$

where $\mathbb{C}_r$ and $\mathbb{C}_t$ are the reference and the test cepstral vector respectively. In a speaker dependent isolated word (alpha-digit vocabulary) recognition trial, the projection measure was shown to be superior to many other distortion measures. The above robust distortion measures were also found to lead to better recognition performance for Lombard speech [34].

3.6 Novel Representations of Speech

The short-time modified coherence (SMC) of speech proposed by Mansour and Juang [34] takes advantage of the inherent coherence in adjacent segments of speech signals to enhance the SNR. It was shown that an unwindowed autocorrelation operation on the impulse response of an all-pole system does not alter its pole structure and estimation of the system parameters may be more reliably accomplished from the autocorrelation function when the signal has been corrupted by noise. The autocorrelation sequence is defined in [34] by

$$\rho_i = \frac{1}{N} \sum_{j=0}^{N-1} x(j)x(j+i) \, , \quad i = 0, 1, \ldots, N$$

which for quasi-stationary signals like speech represents the coherence of adjacent signal segments. All-pole modeling of the autocorrelation sequence results in a more robust signal representation than that of the signal itself. The SNR improvement is data dependent and was found to be around 10 ~ 12 dB for typical SNR's between 0 and 20 dB. For noisy speech recognition at 10 dB SNR, the SMC maintains an accuracy of 98.3% for a speaker-dependent digit test, while the traditional all-pole spectrum representation suffers a severe degradation with accuracy dropping to 39.8%, from 99.2% in clean conditions.

The human auditory system seems to perceive speech better than any machine processor when noise is present. Based on this premise, Ghitza [36] proposed a computational model, called the Ensemble Interval Histogram (EIH), to represent the auditory-nerve firing pattern which might be more robust to noise corruption. The EIH model encompasses 1) use of 85 simulated cochlear filters which split the speech signal in the frequency band of 200-3200 Hz, 2) measurements of level crossing for each of these cochlear filter outputs, and 3) accumulation of the histograms for the level-crossing intervals. The resultant histogram ensemble is reminiscent of a spectrum, but with built-in nonlinearlities and non-uniform frequency resolution that are characteristic of the human auditory processing. When applied to noisy speech recognition in conjunction with a traditional DTW scheme, the EIH was reported to yield significant accuracy improvements for male speech [36].

4.0 Conclusion

Many approaches and algorithms have recently been proposed to cope with the ambient noise, distortion and other noise-induced problems in speech recognition. Proper adoption of these methods proves beneficial when dealing with noisy recognition problems.

REFERENCES

1. B.A. Dautrich et. al., *IEEE ASSP-31*, pp. 793-806, 1983.
2. I. Lecomte et. al., *ICASSP-89*, pp. 512-515.
3. D.B. Roe, *ICASSP-87*, pp. 1139-1142.
4. G.A. Powell et. al., *ICASSP-87*, pp. 173-176.
5. M.B. Carey et. al., *AT&T Bell Lab. Tech. J.*, 63(9), 1984.
6. A. Acero and R.M. Stern, *ICASSP-90*, pp. 849-952.
7. D.B. Pisoni et. al., *ICASSP-85*, pp. 1581-1584.
8. J.-C. Junqua et. al., *ICASSP-90*, pp. 841-844.
9. B. Stanton et. al., *ICASSP-88*, pp. 331-334.
10. R.P. Lippmann et. al., *ICASSP-87*, pp. 705-708.
11. B. Widrow et. al., *Proc. IEEE*, Vol. 63, pp. 1692, 1975.
12. N. Dal Degan et. al., *Signal Processing*, 15, pp. 43-56, 1988.
13. P. Darlington et. al., *ICASSP-85*, pp. 716-719.
14. S.F. Boll, *IEEE ASSP-27*, 2, 113-120, 1979.
15. J.E. Porter et. al., *ICASSP-84*, 18A.2.1-18A.2.4.
16. B.H. Juang et. al., *ICASSP-87*, pp. 2368-2371.
17. Y. Ephraim et. al., *ICASSP-87*, pp. 1324-1327.
18. Y. Ephraim et. al., *IEEE ASSP-32*, 6, 1109-1121.
19. Y. Ephraim et. al., *IEEE ASSP-37*, 12, 1846-1856.
20. V. Viswanathan et. al., *ICASSP-86*, pp. 85-88.
21. D.H. Klatt, *ICASSP-76*, pp. 573-576.
22. J.S. Bridle et. al., Inst. of Acoust., Aut. Conf., Nov. 1984.
23. J.N. Holmes et. al., *ICASSP-86*, pp. 741-744.
24. D. Van Compernolle, *Computer Speech and Language*, vol. 3, pp. 151-167, 1989.
25. A. Nadas et. al., *ICASSP-88*, pp. 517-520.
26. Y. Chen, *ICASSP-87*, pp. 717-720.
27. M. Sugiyama et. al., *Elec. & Comm. Japan*, 65A, 12, 1982.
28. H. Matsumoto et. al., *ICASSP-86*, pp. 769-772.
29. F.K. Soong et. al., *ICASSP-87*, pp. 1257-1230.
30. B.H. Juang et. al., *ICASSP-86*, pp. 765-768.
31. B.A. Hanson et. al., *ICASSP-86*, pp. 757-760.
32. F. Itakura et. al., *ICASSP-87*, pp. 1257-1260.
33. D. Mansour and B. H. Juang, *ICASSP-88*, pp. 36-39.
34. J.-C. Junqua et. al., *ICASSP-89*, pp. 476-479.
35. D. Mansour et. al., *IEEE ASSP-37*, 6, 795-804, 1989.
36. O. Ghitza, *Com. Speech & Lang.*, 1, 2, 109-130, 1986.