We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

11

Speech Recognition in Unknown Noisy Conditions

Ji Ming[1] and Baochun Hou[2]
[1]*Queen's University Belfast*
[2]*University of Hertfordshire*
United Kingdom

1. Introduction

This chapter describes our recent advances in automatic speech recognition, with a focus on improving the robustness against environmental noise. In particular, we investigate a new approach for performing recognition using noisy speech samples without assuming prior information about the noise. The research is motivated in part by the increasing deployment of speech recognition technologies on handheld devices or the Internet. Due to the mobile nature of such systems, the acoustic environments and hence the noise sources can be highly time-varying and potentially unknown. This raises the requirement for noise robustness in the absence of information about the noise. Traditional approaches for noisy speech recognition include noise filtering or noise compensation. Noise filtering aims to remove the noise from the speech signal. Typical techniques include spectral subtraction (Boll, 1979), Wiener filtering (Macho et al., 2002) and RASTA filtering (Hermansky & Morgan, 1994), each assuming *a priori* knowledge of the noise spectra. Noise compensation aims to construct a new acoustic model to match the noisy environment thereby reducing the mismatch between the training and testing data. Typical approaches include parallel model combination (PMC) (Gales & Young, 1993), multicondition training (Lippmann et al., 1987; Pearce & Hirsch, 2000), and SPLICE (Deng et al., 2001). PMC composes a noisy acoustic model from a clean model by incorporating a statistical model of the noise; multicondition training constructs acoustic models suitable for a number of noisy environments through the use of training data from each of the environments; SPLICE improves noise robustness by assuming that stereo training data exist for estimating the corruption characteristics. More recent studies are focused on the approaches requiring less information about the noise, since this information can be difficult to obtain in mobile environments subject to time-varying, unpredictable noise. For example, recent studies on missing-feature theory suggest that, when knowledge of the noise is insufficient for cleaning up the speech features, one may alternatively ignore the severely corrupted features and focus the recognition only on the features with little or no contamination. This can effectively reduce the influence of noise while requiring less knowledge than usually needed for noise filtering or compensation (e.g., Lippmann & Carlson, 1997; Raj et al., 1998; Cooke et al., 2001; Ming et al., 2002). However, missing-feature theory is only effective given partial feature corruption, i.e., the noise only affects part of the speech representation and the remaining part not

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

severely affected by noise can thus be exploited for recognition. This assumption is not realistic for many real-world applications in which the noise will affect all time-frequency components of the speech signal, i.e., we face a full feature corruption problem.

In this chapter, we investigate speech recognition in noisy environments assuming a highly unfavourable scenario: an accurate estimation of the nature and characteristics of the noise is difficult, if not impossible. As such, traditional techniques for noise removal or compensation, which usually assume a prior knowledge of the noise, become inapplicable. We describe a new noise compensation approach, namely *universal compensation*, as a solution to the problem. The new approach combines subband modeling, multicondition model training and missing-feature theory as a means of minimizing the requirement for the information of the noise, while allowing any corruption type, including full feature corruption, to be modelled. Subband features are used instead of conventional fullband features to isolate noisy frequency bands from usable frequency bands; multicondition training provides compensations for expected or generic noise; and missing-feature theory is applied to deal with the remaining training and testing mismatch, by ignoring the mismatched subbands from scoring.

The rest of the chapter is organized as follows. Section 2 introduces the universal compensation approach and the algorithms for incorporating the approach into a hidden Markov model for speech recognition. Section 3 describes experimental evaluation on the Aurora 2 and 3 tasks for speech recognition involving a variety of simulated and realistic noises, including new noise types not seen in the original databases. Section 4 presents a summary along with the on-going work for further developing the technique.

2. Universal Compensation

2.1 The model

Let Φ_0 denote the training data set, containing *clean* speech data, and let $p(X|s, \Phi_0)$ represent the likelihood function of frame feature vector X associated with speech state s trained on data set Φ_0. In this study, we assume that each frame vector X consists of N subband features: $X=(x_1, x_2, ..., x_N)$, where x_n represents the feature for the n'th subband. We obtain X by dividing the whole speech frequency-band into N subbands, and then calculating the feature coefficients for each subband independently of the other subbands. Two different methods have been used to create the subband features. The first method produces the subband MFCC (Mel-frequency cepstral coefficients), obtained by first grouping the Mel-warped filter bank uniformly into subbands, and then performing a separate DCT (discrete cosine transformation) within each subband to obtain the MFCC for that subband (Ming et al., 2002). It is assumed that the separation of the DCT among the subbands helps to prevent the effect of a band-limited noise from being spread over the entire feature vector, as usually occurs within the traditional fullband MFCC. The second method uses the decorrelated log filter-bank energies as the subband features, which are obtained by filtering the log filter-bank energies using a high-pass filter (Ming, 2006). The subband feature framework allows the isolation of noisy bands and selection of the optimal subbands for recognition, thereby improving the robustness against band-selective noise.

The universal compensation approach comprises two steps. The first step is to simulate the effect of noise corruption. This is done by adding noise into the clean training data Φ_0. We have primarily added white noise at variable signal-to-noise ratios (SNRs) to simulate the variation of noise, but different types of noises could be used depending on the expected

environments. Assume that this leads to multiple training sets Φ_0, Φ_1, …, Φ_K, where Φ_k denotes the k'th training set derived from Φ_0 with the addition of a specific level of corruption. Then a new likelihood function for the test frame vector can be formed by combining the likelihood functions trained on the individual training sets:

$$p(X \mid s) = \sum_{k=0}^{K} p(X \mid s, \Phi_k) P(\Phi_k \mid s) \tag{1}$$

where $p(X \mid s, \Phi_k)$ is the likelihood function of frame vector X associated with state s trained on data set Φ_k, and $P(\Phi_k \mid s)$ is the prior probability for the occurrence of the corruption condition Φ_k at state s. Eq. (1) is a multicondition model. A recognition system based on Eq. (1) should have improved robustness to the noise conditions seen in the training sets $\{\Phi_k\}$, as compared to a system based on $p(X \mid s, \Phi_0)$.

The second step of the approach is to make Eq. (1) robust to noise conditions not fully matched by the training sets $\{\Phi_k\}$ without assuming extra information about the noise. One way to achieve this is to ignore the heavily mismatched subbands and focus the score only on the matching subbands. Let $X=(x_1, x_2, ..., x_N)$ be a test frame vector and X_k be a specific subset of features in X which are corrupted at noise condition Φ_k. Then, using X_k in place of X as the test vector for each training noise condition Φ_k, Eq. (1) can be modified as

$$p(X \mid s) = \sum_{k=0}^{K} p(X_k \mid s, \Phi_k) P(\Phi_k \mid s) \tag{2}$$

where $p(X_k \mid s, \Phi_k)$ is the marginal likelihood of the matching feature subset X_k, derived from $p(X \mid s, \Phi_k)$ with the mismatched subband features ignored to improve mismatch robustness between the test frame X and the training noise condition Φ_k. For simplicity, assume independence between the subband features. So the marginal likelihood $p(X_{sub} \mid s, \Phi_k)$ for any subset X_{sub} in X can be written as

$$p(X_{sub} \mid s, \Phi_k) = \prod_{x_n \in X_{sub}} p(x_n \mid s, \Phi_k) \tag{3}$$

where $p(x_n \mid s, \Phi_k)$ is the likelihood function of the n'th subband feature at state s trained under noise condition Φ_k.

Multicondition or multi-style model training (e.g., Eq. (1)) has been a common method used in speech recognition to account for varying noise sources or speaking styles. The universal compensation model expressed in Eq. (2) is novel in that it combines multicondition model training with missing-feature theory, to ignore noise variations outside the given training conditions. This combination makes it possible to account for a wide variety of testing conditions based on limited training conditions (i.e., Φ_0 through Φ_K), as will be demonstrated later in the experiments.

Missing-feature theory is applied in Eq. (2) for ignoring the mismatched subbands. However, it should be noted that the approach in Eq. (2) extends beyond traditional missing-feature approaches. Traditional approaches assess the usability of a feature against its clean data, while the new approach assesses this against the data containing variable degrees of corruption, modelled by the different training conditions Φ_0 through Φ_K. This allows the model to use noisy features, close to or matched by the noisy training conditions,

for recognition. These noisy features, however, may become less usable or unusable with traditional missing-feature approaches due to their mismatch against the clean data.

Given a test frame X, the matching feature subset X_k for each training noise Φ_K may be defined as the subset in X that gains maximum likelihood over the appropriate noise condition. Such an estimate for X_k is not directly obtainable from Eq. (3). This is because the values of $p(X_{sub}|s, \Phi_k)$ for different sized subsets X_{sub} are of a different order of magnitude and are thus not directly comparable. One way around this is to select the matching feature subset X_k for noise condition Φ_k that produces maximum likelihood ratio for noise condition Φ_k as compared to all other noise conditions $\Phi_j \neq \Phi_k$. This effectively leads to a posterior probability formulation of Eq. (2). Define the posterior probability of state s and noise condition Φ_k given test subset X_{sub} as

$$P(s, \Phi_k | X_{sub}) = \frac{p(X_{sub} | s, \Phi_k) P(s, \Phi_k)}{\sum_{\varsigma, j} p(X_{sub} | \varsigma, \Phi_j) P(\varsigma, \Phi_j)} \tag{4}$$

On the right, Eq. (4) performs a normalization for $p(X_{sub}|s, \Phi_k)$ using the average likelihood of subset X_{sub} calculated over all states and training noise conditions, with $P(s, \Phi_k) = P(\Phi_k|s)P(s)$ being a prior probability of state s and noise condition Φ_k. The normalization makes it possible to compare the probabilities associated with different feature subsets X_{sub} and to obtain an estimate for X_k based on the comparison. Specifically, we can obtain an estimate for X_k by maximizing the posterior probability $P(s, \Phi_k|X_{sub})$ with respect to X_{sub}. Dividing the numerator and denominator of Eq. (4) by $p(X_{sub}|s, \Phi_k)$ gives

$$P(s, \Phi_k | X_{sub}) = \frac{P(s, \Phi_k)}{P(s, \Phi_k) + \sum_{\varsigma, j \neq s, k} P(\varsigma, \Phi_j) p(X_{sub} | \varsigma, \Phi_j) / p(X_{sub} | s, \Phi_k)} \tag{5}$$

Therefore maximizing posterior probability $P(s, \Phi_k|X_{sub})$ with respect to X_{sub} is equivalent to the maximization of likelihood ratios $p(X_{sub}|s, \Phi_k)/p(X_{sub}|\varsigma, \Phi_j)$, for all $(\varsigma, \Phi_j) \neq (s, \Phi_k)$, by choosing X_{sub}. The universal compensation model, Eq. (2), can be expressed in terms of the posterior probabilities $P(s, \Phi_k|X_{sub})$ as follows (the expression will be derived later)

$$p(X | s) \propto \sum_{k=0}^{K} \max_{X_{sub} \subset X} p(s, \Phi_k | X_{sub}) \tag{6}$$

where the maximization at each noise condition Φ_k accounts for the selection of the optimal set of subband features for that noise condition.

2.2 Incorporation into a hidden Markov model (HMM)

Assume a speech signal represented by a time sequence of T frames $X_{1\sim T}$=(X(1), X(2), …, X(T)), and assume that the signal is modelled by an HMM with parameter set λ. Based on

the HMM formulation, the likelihood function of $X_{1\sim T}$, given the state sequence $S_{1\sim T}$=(s(1), s(2), …, s(T)), where s(t) is the state for frame X(t), can be written as

$$p(X_{1\sim T} \mid S_{1\sim T}, \lambda) = \prod_{t=1}^{T} p(X(t) \mid s(t)) \tag{7}$$

where $p(X \mid s)$ is the state-based observation probability density function with the HMM. To incorporate the above universal compensation approach into the HMM, we first express the state-based observation density $p(X \mid s)$ in terms of $P(s, \Phi_k \mid X)$, i.e., the posterior probabilities of state s and noise condition Φ_k given frame vector X. Using Bayes's rules it follows

$$\begin{aligned} p(X \mid s) &= \frac{P(s \mid X)p(X)}{P(s)} \\ &= \frac{\sum_{k=0}^{K} P(s, \Phi_k \mid X)}{P(s)} p(X) \end{aligned} \tag{8}$$

The last term in Eq. (8), p(X), is not a function of the state index and thus has no effect in recognition. Substituting Eq. (8) into Eq. (7), replacing each $P(s, \Phi_k \mid X)$ with the maximized posterior probability for selecting the optimal set of subbands and assuming an equal prior probability P(s) for all the states, we obtain a modified HMM which incorporates the universal compensation approach

$$p(X_{1\sim T} \mid S_{1\sim T}, \lambda) \propto \prod_{t=1}^{T} \sum_{k=0}^{K} \max_{X_{sub} \subset X(t)} P(s(t), \Phi_k \mid X_{sub}) \tag{9}$$

where $P(s, \Phi_k \mid X_{sub})$ is defined in Eq. (4) with $P(s, \Phi_k)$ replaced by $P(\Phi_k \mid s)$ due to the assumption of a uniform prior P(s). In our experiments, we further assume a uniform prior $P(\Phi_k \mid s)$ for noise conditions Φ_k, to account for the lack of prior knowledge about the noise.

2.3 Algorithm for implementation

The search in Eq. (9) for the matching feature subset can be computationally expensive for frame vectors with a large number of subbands (i.e., N). We can simplify the computation by approximating each $p(X_{sub} \mid s, \Phi_k)$ in Eq. (4) using the probability for the union of all subsets of the same size as X_{sub}. As such, $p(X_{sub} \mid s, \Phi_k)$ can be written, with the size of X_{sub} indicated in brackets, as (Ming et al. 2002)

$$p(X_{sub}(M) \mid s, \Phi_k) \propto \sum_{\text{all } X'_{sub}(M) \subset X} p(X'_{sub}(M) \mid s, \Phi_k) \tag{10}$$

where $X_{sub}(M)$ represents a subset in X with M subband features ($M \leq N$). Since the sum in Eq. (10) includes all feature subsets, it includes the matching feature subset that can be assumed to dominate the sum due to the best data-model match. Therefore Eq. (4) can be rewritten, by replacing $p(X_{sub} \mid s, \Phi_k)$ with $p(X_{sub}(M) \mid s, \Phi_k)$, as

$$P(s,\Phi_k \mid X_{sub}(M)) = \frac{p(X_{sub}(M) \mid s,\Phi_k)P(s,\Phi_k)}{\sum_{\varsigma,j} p(X_{sub}(M) \mid \varsigma,\Phi_j)P(\varsigma,\Phi_j)} \tag{11}$$

Note that Eq. (11) is not a function of the identity of X_{sub} but only a function of the size of X_{sub} (i.e., M). Using $P(s, \Phi_k | X_{sub}(M))$ in place of $P(s, \Phi_k | X_{sub})$ in Eq. (9), we therefore effectively turn the maximization for the exact matching feature subset, of a complexity of $O(2^N)$, to the maximization for the size of the matching feature subset, with a lower complexity of O(N). The sum in Eq. (10) over all $p(X_{sub}(M) | s, \Phi_k)$ for a given number of M features, for $0<M\leq N$, can be computed efficiently using a recursive algorithm assuming independence between the subbands (i.e., Eq. (3)). We call Eq. (11) the *posterior union model*, which has been studied previously (e.g., Ming et al., 2006) as a missing-feature approach without requiring identity of the noisy data. The universal compensation model Eq. (9) is reduced to a posterior union model with single, clean condition training (i.e., K=0).

3. Experimental Evaluation

The following describes the experimental evaluation of the universal compensation model on the Aurora 2 and 3 databases, involving a variety of simulated and realistic noises, including additional noise types not seen in the original databases. In all the experiments, the universal compensation system assumed no prior information about the noise.

3.1 Experiments on Aurora 2

Aurora 2 (Pearce & Hirsch, 2000) is designed for speaker independent recognition of digit sequences in noisy conditions. Aurora 2 involves nine different environments (eight noisy and one noise-free) and two different channel characteristics. The eight environmental noises include: subway, bubble, car, exhibition hall, restaurant, street, airport and train station. The two channel characteristics are G712 and MIRS. Aurora 2 has been divided into three test sets, each corresponding to a different set of noise conditions and/or channel characteristics. These are: 1) test set A - including clean data and noisy data corrupted by four different noises: subway, babble, car and exhibition hall, each at six different SNRs: 20 dbB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB, filtered with a G712 characteristic; 2) test set B - including data corrupted by four different noises: restaurant, street, airport and train station, each at the same range of SNRs as in test set A, filtered with a G712 characteristic; and 3) test set C - including data corrupted by two different noises: subway and street, each at the same range of SNRs as in test set A, filtered with an MIRS characteristic.

Aurora 2 offers two training sets, for two different training modes: 1) clean training set, consisting of only clean training data filtered with a G712 characteristic; and 2) multicondition training set, consisting of both clean data and multicondition noisy data involving the same four types of noise as in test set A, each at four different SNRs: 20 dB, 15 dB, 10 dB, 5 dB, and filtered with a G712 characteristic - also the same as for test set A. As such, it is usually assumed that the multicondition training set matches test set A more closely than it matches test set B. However, as noted in (Pearce & Hirsch, 2000), the noises in test set A seem to cover the spectral characteristics of the noises in test set B, and therefore no significant differences in performance have been found between test set A and test set B based on the model trained on the multicondition data. Mismatches exist between the

multicondition training data and test set C, because of the different channel characteristics (i.e., G712 versus MIRS).

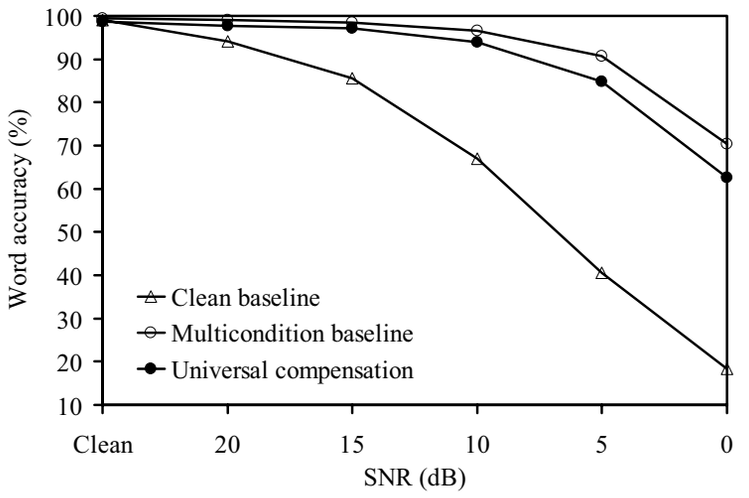


Figure 1. Word accuracy on the Aurora 2 database

The universal compensation model was compared with two baseline systems. The first baseline system was trained on the clean training set and the second was trained on the multicondition training set. The universal compensation model was trained using only the clean training set. This clean training set was expanded by adding computer-generated wide-band noise to each of the training utterances at ten different SNR levels, starting with SNR = 20 dB, reducing 2 dB every level, until SNR = 2 dB. This gives a total of eleven corruption levels (including the no corruption condition) for training the universal compensation model. The wide-band noise used in the training was computer-generated white noise filtered by a low-pass filter with a 3-dB bandwidth of 3.5 kHz. In modelling the digit words, the same HMM topology was adopted for all the three models: each word being modelled with 16 states, and each state being modelled with Gaussian mixture densities. Thirty two mixtures were used in each state for the universal compensation model and the multicondition baseline model, while 3 mixtures were used in each state for the clean baseline model trained only on the clean data. The speech signal, sampled at 8 kHz, was divided into frames of 25 ms at a frame period of 10 ms. The universal compensation model used subband features, consisting of 6 subbands derived from the decorrelated log filter-bank energies, as the feature set for each frame. The baseline systems used fullband MFCC as the feature set. Both models included the first- and second-order time differences as dynamic features. More details of the implementation can be found in (Ming, 2006).

Fig. 1 shows the word accuracy rates for the three systems: clean baseline, multicondition baseline and universal compensation, as a function of SNR averaged over all the noise conditions in test set A, B and C. As indicated in Fig. 1, the universal compensation model significantly improved over the clean baseline model, and achieved an average performance close to that obtained by the multicondition baseline model trained on the Aurora noisy

training data. Note that the universal compensation model achieved this based only on the clean training data and simulated noisy training data, without having assumed any knowledge about the actual test noise.

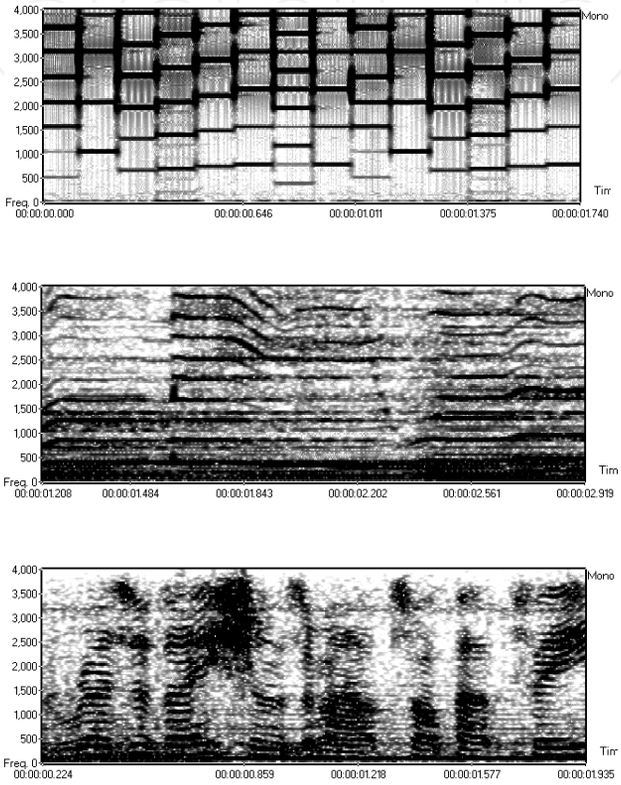


Figure 2. Spectrograms of three new noises unseen in Aurora 2. From top to bottom: mobile phone ring, pop song, broadcast news

To further investigate the capability of the universal compensation model to offer robustness for a wide variety of noises, three new noise conditions unseen in the Aurora 2 database were added in the test. The three new noises are: 1) a polyphonic mobile phone ring, 2) a pop song segment with a mixture of background music and the voice of a female singer, and 3) a broadcast news segment from a male speaker. Fig. 2 shows the spectral characteristics of the three new noises. Fig. 3 shows a comparison between the universal compensation model and the multicondition baseline model across the Aurora 2 noise conditions and the new noise conditions. As expected, the multicondition baseline trained using the Aurora data performed better than the universal compensation model under the Aurora 2 noise conditions. However, the multicondition baseline performed poorer than the universal compensation model for all the unseen noises, due to the mismatched conditions between the training and testing. The universal compensation model achieved a better

average performance across all the noise conditions, indicating improved robustness for dealing with unknown/mismatched noises.

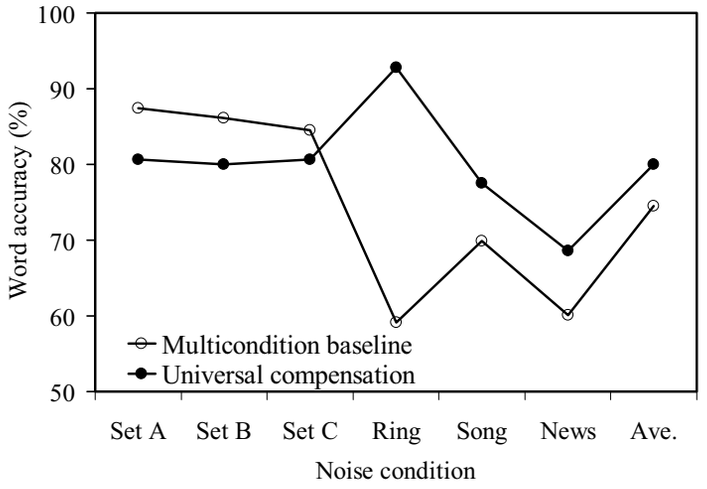


Figure 3. Word accuracy in different noise conditions within and outside the Aurora 2 database (averaged over SNRs between 0-10 dB)

The universal compensation approach involves a combination of multicondition model training and missing-feature theory. The importance of the combination, in terms of improved recognition performance, is studied. We first considered a system which was built on the same simulated noisy training data as used for the universal compensation model, but did not apply missing-feature theory to optimally select the subbands for scoring. The system thus used the full set of subbands for recognition. Comparisons were conducted for all the Aurora 2 noises and the three new noises as described above. Fig. 4 shows the absolute improvement in word accuracy obtained by the universal compensation model over the system without optimal subband selection. The results indicate that the optimal subband selection in the universal compensation model has led to improved accuracy in all tested noisy conditions. As expected, the improvement is more significant for those noises with a spectral structure significantly different from the wide-band noise spectral structure as used in the universal compensation model for noise compensation. In our experiments, these noises include, for example, the mobile phone ring, pop song, broadcast news and airport noises. Fig. 4 also indicates that the absolute improvement from the optimal subband selection is more significant in low SNR conditions (except for the exhibition-hall noise).

The above experimental results indicate the importance of the missing-feature component in the universal compensation model, for achieving robustness to mismatched training and testing. Likewise, the multicondition training component in the model plays an equally important role, particularly for dealing with broad-band noise corruption for which the conventional missing-feature methods usually fail to function. To show this, we considered a system which performed optimal subband selection as the universal compensation model, but was not trained using the simulated multicondition noisy data. Rather, it was trained using only the clean training data. Comparisons were conducted on test set A of the Aurora 2 database. Fig. 5 shows the absolute improvement in word accuracy obtained by the

universal compensation model over the system with the missing-feature component but without being trained on the multicondition data. This missing-feature system performed better than the clean baseline model (i.e., the baseline model trained on the clean training data), due to the optimal selection of the subbands, but worse than the universal compensation model. The broad-band nature of the noises in test set A causes the poor performance for this missing-feature system.

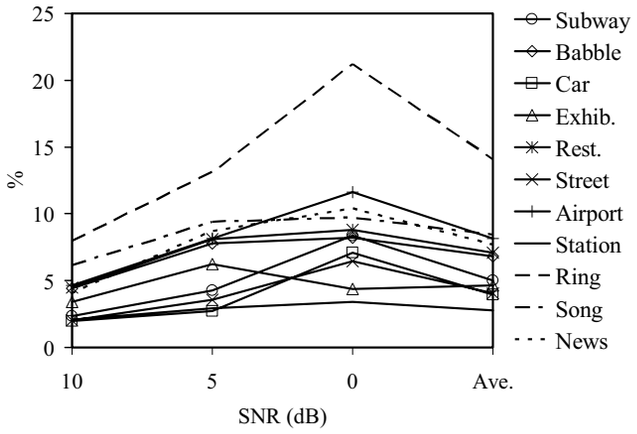


Figure 4. Absolute improvement in word accuracy obtained by optimal subband selection

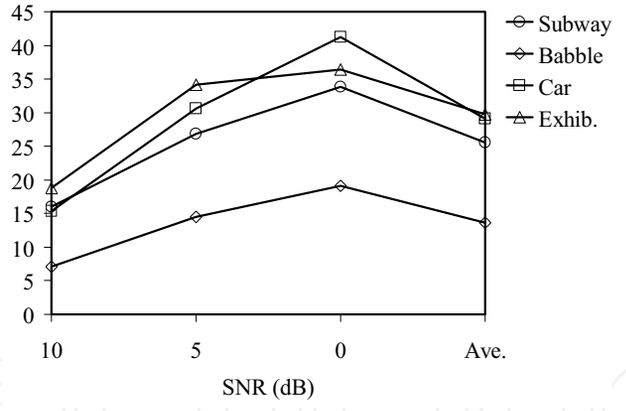


Figure 5. Absolute improvement in word accuracy obtained by multicondition training

3.2 Experiments on Aurora 3

Unlike Aurora 2, the Aurora 3 database consists of digit sequences (in four languages – Danish, Finnish, German and Spanish) recorded in real-world in-car environments, with realistic noise and channel effects. Speech data were recorded in three different noisy (driving) conditions - quite, low noise and high noise, and each utterance was recorded simultaneously by using two microphones, a close-talk microphone and a hand-free microphone. Three experimental conditions are defined in Aurora 3: 1) well-matched

condition in which the training and testing data sets contain well-matched data for both the microphones and noise conditions; 2) moderately-mismatched condition in which the training and testing data are both from the hand-free microphone but differ in noise levels - quite and low-noise data for training and high-noise data for testing; 3) highly-mismatched condition in which the training and testing sets differ in both the microphone and noise levels - the data collected using the close-talk microphone in all the three conditions are used for training and the data collected using the hand-free microphone in the low-noise and high-noise conditions are used for testing. The hand-free microphone picked up more noise than the close-talk microphone from the background. In our experiments, the universal compensation model was trained using the training data for the highly-mismatched condition, by treating the close-talk data as "clean" data. The close-talk training data were expended by adding simulated wide-band noise at ten different SNRs between 2 – 20 dB. These simulated noisy speech data were used to train the universal compensation model, which used the same subband feature structure as for Aurora 2.

Training vs. Testing	Danish	Finnish	German	Spanish	Average
Well matched	12.7	7.3	8.8	7.1	8.9
Moderately mismatched	32.7	19.5	18.9	16.7	21.9
Highly mismatched	60.6	59.5	26.8	48.5	48.9
Average	35.3	28.8	18.2	24.1	26.6

Table 1. Word error rates on the Aurora 3 database, by the ETSI baseline system

Training vs. Testing	Danish	Finnish	German	Spanish	Average
Well matched	11.2	6.1	7.5	6.7	7.9
Moderately mismatched	26.8	17.2	16.3	15.5	18.9
Highly mismatched	19.9	12.5	13.7	12.2	14.6
Average	19.3	11.9	12.5	11.5	13.8

Table 2. Word error rates on the Aurora 3 database, by the universal compensation model

Table 1 shows the word error rates produced by the ETSI (European Telecommunications Standards Institute) baseline system, included for comparison. Table 2 shows the word error rates produced by the universal compensation model. As indicated in Tables 1 and 2, the universal compensation model performed equally well as the baseline system trained and tested in the well-matched conditions. The universal compensation model outperformed the baseline system when there were mismatches between the training and testing conditions. The average error reduction is 70.1%, 13.7% and 11.2%, respectively, for the highly-mismatched, moderately-mismatched and well-matched conditions.

4. Conclusion

This chapter investigated the problem of speech recognition in noisy conditions assuming absence of prior information about the noise. A method, namely universal compensation, was described, which combines multicondition model training and missing-feature theory to model noises with unknown temporal-spectral characteristics. Multicondition training can be conducted using simulated noisy data, to provide a coarse compensation for the noise, and missing-feature theory is applied to refine the compensation by ignoring noise

variations outside the given training conditions, thereby accommodating mismatches between the training and testing conditions. Experiments on the noisy speech databases Aurora 2 and 3 were described. The results demonstrate that the new approach offered improved robustness over baseline systems without assuming knowledge about the noise. Currently we are considering an extension of the principle of universal compensation to model new forms of signal distortion, e.g., handset variability, room reverberation, and distant/moving speaking. To make the task tractable, these factors can be "quantized" as we did for the SNR. Missing-feature approaches will be used to deemphasize the mismatches while exploiting the matches arising from the quantized data.

5. References

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 27, Apr 1979, pp. 113-120.

Cooke, M.; Green, P.; Josifovski, L. & Vizinho, A. (2001) Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, Vol. 34, 2001, pp. 267-285.

Deng, L.; Acero, A.; Jiang, L.; Droppo, J. & Hunag, X.-D. (2001). High-perfromance robust speech recognition using stereo training data, *Proceedings of ICASSP*, pp. 301-304, Salt Lake City, Utah, USA, 2001.

Gales, M. J. F. & Young, S. (1993). HMM recognition in noise using parallel model combination, *Proceedings of Eurospeech'93*, pp. 837-840, Berlin, Germany, 1993.

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, Vol. 4, Oct 1994, pp. 578-589.

Lippmann, R. P.; Martin, E. A. & Paul, D. B. (1987). Multi-style training for robust isolated-word speech recognition, *Proceedings of ICASSP*, pp. 705-708, Dallas, TX, USA, 1987.

Lippmann, R. P. & Carlson, B. A. (1997). Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise, *Proceedings of Eurospeech*, pp. 37-40, Rhodes, Greece, 1997.

Macho, D.; Mauuary, L.; Noe, B.; Cheng, Y. M.; Ealey, D.; Jouver, D.; Kelleher, H.; Pearce, D. & Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases, *Proceedings of ICSLP*, pp. 17-20, Denver, CO, USA, 2002.

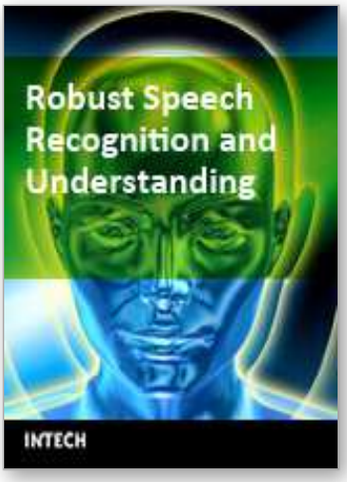
Ming, J.; Jancovic, P. & Smith, F. J. (2002). Robust speech recognition using probabilistic union models. *IEEE Transactions on Speech and Audio Processing*, Vol. 10, Sep 2002, pp. 403-414.

Ming, J.; Lin, J. & Smith, F. J. (2006). A posterior union model with applications to robust speech and speaker recognition. *EURASIP Journal on Applied Signal Processing*, Vol. 2006, 2006, Article ID 75390.

Ming, J. (2006). Noise compensation for speech recognition with arbitrary additive noise. *IEEE Transactions on Speech and Audio Processing*, Vol. 14, May 2006, pp. 833-844.

Pearce, D. & Hirsch, H.-G. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *Proceedings of ISCA ITRW ASR*, Paris, France, 2000.

Raj, B.; Singh, R. & Stern, R. M. (1998). Inference of missing spectrographic features for robust speech recognition, *Proceedings of ICSLP*, pp. 1491-1494, Sydney, Australia, 1998.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen