Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 11-20

# A PRELIMINARY STUDY ABOUT ROBUST SPEECH RECOGNITION FOR A ROBOTICS APPLICATION

*Xue Wang and Louis C. W. Pols*

## Abstract

In this study the specific problem of robustness in automatic speech recognition under various acoustic conditions is reviewed. The currently most used techniques to solve this problem are discussed. One specific technique (RASTA-PLP) is then implemented and compared with a more conventional approach (MFCC) for the frond-end processing for the recogniser. This technique is tested in a fictitious application (with a 26-word vocabulary, finite-state grammar, 10 speakers, and added Gaussian noise) intended for use in an autonomous robotic system. Preliminary word recognition results show promising indications that the chosen technique is suitable for improving the robustness.

## 1. Introduction

In this paper the preliminary results of a pilot study on robust speech recognition are presented. This study serves as a first step towards an intended long-term research project on a sound understanding machine (SUM) for an autonomous robot. If funding is approved, this project will be carried out at the Institute of Phonetic Sciences (IFA) and at the Section of Intelligent Autonomous Systems (IAS) of the University of Amsterdam. The results of the present pilot study will be useful for further development of the SUM project.

In section 2 the general problem of robustness in automatic speech recognition (ASR) is discussed. The three most promising approaches in this field are reviewed. Next the design of the system for the pilot study is given. Finally, the results of the pilot study are presented and future work, related to the SUM project, is discussed.

## 2. Robust speech recognition

Nowadays, a commercial large-vocabulary continuous speech recogniser can already reach a performance level of 95% word accuracy for the clean voice of a single person via a close-talking, head-mounted microphone. However, in many real world applications, speech must be captured via a hands-free and "head-free" microphone. In such situations, ambient noise will certainly interfere. The recognition algorithm for clean speech usually performs very unsatisfactorily for noisy speech. Even for a moderately noisy environment and a small vocabulary task, additional techniques are

required to reach an acceptable recognition rate. The issue of robustness in ASR has mainly been translated into technical approaches to better accommodate into a speech recogniser various sources of variability introduced by a corrupted speech signal. The development in this area currently attracts much research interest (e.g. Juang, 1995), apart from the effort put into the deployment of speech technology.

Much of the variability in speech, including speaker variability, is handled in modern ASR systems by statistical methods. The more complex the type of variability is that has to be accommodated in a recogniser, the more sophisticated the model structures get. To this end we see for instance monophone HMMs (hidden Markov models), triphones, duration models (Wang, 1997) and phonological rules. Various kinds of distortions in the speech signal, as well as noise, introduce more variability compared to the clean speech signal. A robust speech recogniser should then also be able to deal with this new variability, either by the existing structure, or, if necessary, by additional structures.

After several decades of theoretical and technical development in speech science and technology to solve real-life problems, one way to approach the problem is to distinguish between the variable and the invariable parts of the speech signal. In automatic speech recognition, one generally ignores the variable part whereas focusing on the invariable linguistic content carried in the signal. The current level of sophistication, however, is that we do not know whether separating the variable part from the invariable part in speech is the ultimate paradigm for ASR. The intrinsic reason behind this may still be a lack of a sufficient understanding of the human-to-human speech communication process. For lack of a better solution, we might have to use this paradigm still for quite a time. Usually the invariability is closely related to the basic information-carrying units in the signal of speech and various environmental sounds. In other words, we want our system to recognise for example voice commands and a few interested sounds (like slamming doors, breaking glass, telephone ringing etc.), irrespective of the background noise these signals are embedded in.

Based on the well-developed standard system structure of a speech recogniser (based on e.g. HMM), speech and other sounds should be represented within the ASR system by a finite number of measures (or features) of the signal. This is because the entirety of the information in the acoustic signal is too much to process, yet not all kinds of information are relevant for specific tasks. Feature extraction (usually achieved in the "front-end" of an ASR) has been based on some understanding of the human communication processes and on years of engineering experience. One of the approaches for robust ASR is to select such features that are relatively insensitive to the unwanted variations, while keeping the part, which represents the message in the signal relatively intact. This is the "feature-based" approach for robust recognition. Essentially the rest of the system is the same as for the recognition of clean speech. An alternative set-up could be to use a "speech-enhancement" system followed by an ASR, in which the former produces speech with an enhanced quality for the latter. However such approaches have not currently been applied in real world ASR systems, mainly due to the heavy calculation load for the added enhancement algorithms and the limited benefits they currently provide.

Another approach is called "model-based". Here "model" refers to the basic modelling units (e.g. an HMM), instead of the mechanism embedded in the feature-extracting front-end that models some human speech production or perception processes (according to our terminology, the function of the front-end is already included in the feature-based approach). In such a model-based approach all information concerning various kinds of variability is retained after the front-end. The part of the system that models the patterns to be recognised, compensates, de-emphasises, or adapts to the unwanted variability. Because of the strong (statistical)

modelling power of the mathematical framework of HMMs (or neural nets), there is still a potential for these models to do more sophisticated jobs. For instance, one can refine the internal structure of these models, or one can redefine the mapping between these models and particular physical events (e.g. to model the pure sound categories and noises by separate models).

The third approach to the problem of robustness is to use a microphone array. This approach can be considered an enhancement method implemented outside of the recogniser. These three approaches are discussed below in more detail.

### 2.1. Feature-based approach

For clean speech, linear predictive coefficient (LPC) is commonly used, resulting in good recognition performance. Mostly LPC is implemented as an all-pole AR model to capture the vocal-tract properties of vowel-like sounds. Usually autocorrelation methods are used on a windowed section of the speech samples. When LPC is applied directly to a signal that contains distortions introduced by e.g. communication channel and environment, the performance degrades.

The currently most pronounced feature-based approach for robust speech recognition is PLP (perceptual linear predictive, Hermansky, 1990), sometimes combined with RASTA (relative spectrum, Hermansky et al., 1991 and Hermansky and Morgan, 1994). In PLP analysis, additional steps to LPC are included to account for the human auditory system:

(1) A critical-band spectral resolution mechanism integrates the frequency components within a total of 15-18 frequency bands based on inner-ear analogy (nearly logarithmic or Bark spacing of the centre frequencies and the asymmetrical shape of each band);
(2) An equal-loudness curve to approximate the nonequal sensitivity of the human hearing in different frequency regions (e.g. flat between 400 and 1200 Hz, 6 dB/oct between 1200 and 3100 Hz);
(3) An intensity-to-loudness conversion approximated by a power of (1/3).

In practice, the implementation of various steps in the PLP processing is also based on computational efficiency, and performed in either the time or the frequency domain.

In general, PLP representation of speech mimics some aspects of the human auditory perception mechanism, capturing useful speech information while de-emphasising e.g. speaker variability. Experiments also showed that PLP outperformed LP in ASR. There are evidences that the human auditory system captures speech in an efficient way (its best ability is dynamically tuned to the important speech feature). In this respect, if the target of our recogniser is speech, PLP can produce a better feature than LP. In principle, other computer algorithms may be developed which do not mimic humans, but are more efficient for representing speech. However, such algorithms have not yet been developed.

RASTA, on the other hand, usually consists of additional processing steps to PLP. The main additional steps are applied after the critical-band integration of PLP. Based on an observation that human perception is more sensitive to relative changes, the static or slow-changing part of the signal (in each critical-band) is effectively filtered out. This is a band-pass filtering process (which emphasis signals around 4 Hz) applied on the already processed time sequence of the feature vectors (short-term spectra, usually at a 100 Hz rate).

Due to the fact that most distortions due to unknown channel characteristics (telephone lines or microphones) are effectively convolutive, RASTA is best implemented in a logarithmic domain of the spectrum, which makes the different parts additive. Non-linear conversions are required. RASTA-PLP outperforms PLP for recognition of channel-distorted speech.

### 2.2. Model-based approach

There also exists processing power after the front-end processing in a recogniser. In the basic structure of an HMM-based recogniser, the states of the HMMs collect the statistics of the various patterns in the speech signal. One of the commonly used parametrical distributions of these states is "Gaussian mixtures".

One of the model-based approaches for robustness of ASR is to separately model the different parts of speech by different parts of the model structures. For example, separate sets of HMM parameters (e.g. the mean and variance of the Gaussians) are defined for the useful information and the distortion part, respectively. To this end equalisation methods (Juang and Paliwal, 1992) and adaptation methods (Gales and Woodland, 1996) have been developed.

Though not necessary by definition, these model-based approaches are mostly designed to treat the distortions in a paradigm of adaptation. So the ASR system is firstly trained for clean speech or speech with known environment properties, and the ASR system is capable of adapting to the properties of speech of unknown or new environments. Sometimes the problem is addressed in a framework of a "mismatch" between training and testing conditions for the statistical ASR (Sankar and Lee, 1996).

The model-based approach is not yet considered in the current study. It may be considered in the long-term SUM project.

### 2.3. Microphone-array approach

One point, common to both the feature-based and the model-based approaches, is that additional structures (regularities) in the signal are found and built into the ASR system to take advantage of the *difference* between the wanted speech and the distorting signals. Similar to this, another approach has been developed which makes use of an array of microphones to capture the acoustic signal. In this case, the additional structure that tries to separate the speech and distorting signals is physically located outside the recogniser. The geometrical relations between the different sound sources will result in different signals arriving at the microphones in the array. Of course, special algorithms have to be developed to actually make use of the extra information obtained from the microphone-array (Oh and Viswanathan, 1995). In principle, a recogniser used with a microphone array should also be tailored to the kind of multi-channel signal in order to use the additional information optimally.

The difference in signals via the different microphones can also be used to locate a sound source. Since this is also a task of (a group of) voice-interactive robots, we can hope to ultimately use this information to do both jobs. Probably still more information will be needed for both sound localisation and enhancement of the sound quality, such as the information obtained at different points in time (time of arrival) from a moving sound source. The algorithm will be rather complicated due to this double-task. There must be a trade off between the complexity of the system and the task requirement.

## 3. First implementation of a feature-based approach: a pilot study

As a first step in the SUM (sound understanding machine) project, a pilot study has been carried out. We tried to define a sub-task from the research and development of the main SUM project, that should be self-contained. It also should provide useful results for the SUM project. Judging from the scheduling of the SUM project, at this point in time when the robot system is not yet implemented, the starting period of the project can concentrate on the robust recognition problem. This problem should in any case be solved for the SUM project, because the robot, on which the SUM is mounted, is going to act in a noisy real-world environment. Another reason for the definition of this specific short project stems from the extensive experience of the author and the IFA in speech recognition in general, and from a recent shift of research interest into robust speech recognition of the first author, in particular.

Analysing the three approaches reviewed above, we have chosen to investigate the feature-based approach first in the short project. This approach can be implemented in the current software framework of the recognition system used until now at the IFA. The second model-based approach will be investigated after completing this short project, probably using also the same framework. The third (microphone array) approach will be implemented later, but this is only possible when the hardware and overall set-up of the robot system are available. Furthermore, the main task of the microphone array will be sound source localisation, rather than speech enhancement. However, since the microphone array will be available in the SUM project, the extra information can be helpful also for enhancing the quality of the speech signal arriving at the recogniser, thus possibly improving the recognition performance.

### 3.1. Recognition system and algorithms

The actual (non real-time) recognition system is based on a software development tool-kit called HTK (HMM Tool Kit, Young, 1992), developed at the Cambridge University. All the source codes (in C) are available, therefore it is possible to modify the functionality. The major modification of HTK concerns the signal processing front-end which converts the speech samples into features, given in the form of *analysis frames*. The original HTK supports, among others, MFCC (Mel Frequency Cepstrum Coefficients). We decided to add RASTA-PLP into HTK, so that it can be easily compared with the MFCC, whereas the rest of the recogniser will be the same. Several versions of RASTA and/or PLP processing algorithms (Hermansky et al., 1991) were integrated into the HTK.

The recognition system design for this short project is as follows. Since the total number of the words to be recognised is small (26, see section 3.2 below), we define HMMs on whole-words instead of any sub-word units. Based on our previous experience in duration modelling (Wang, 1997), all HMMs have a linear transition topology. We considered 4 states to be sufficient to model the number of different portions within each of the words. The observation density of each of these states is a mixture of 8 Gaussian components. The observation vector consists of 12 basic coefficients, the frame energy, and their first and second time-derivatives, together making a 39-dimensional vector. The basic coefficients are either MFCC or RASTA-PLP.

### 3.2. Recognition task and speech database

The recognition task can be best shown by the finite-state grammar used in the system, as given in the block below. It consists of a total of 26 Dutch words, to be used for a fictitious robot task in which the robot is ordered by voice commands to come, go away, go to certain directions and locations, or fetch coffee or tea.

```
$getal = een | twee | drie | vier | vijf | zes | zeven | acht |
 negen | nul;
$num = nummer [Ag%%] $getal [Ag%%] $getal;
$richting = links | rechts;
$voor_achter = vooruit | achteruit;
$drank = koffie | thee;
$Telefoon1 = Tele1 [ [Ag%%] Tele1 ] [ [Ag%%] Tele1 ];
$Telefoon2 = Tele2 [ [Ag%%] Tele2 ] [ [Ag%%] Tele2 ];
$geluid = Deur | Motor | $Telefoon1 | $Telefoon2;
$cmd =      kom [[Ag%%] hier] |
      [[ga [Ag%%]] naar [Ag%%]] $richting |
      [ga [Ag%%]] $voor_achter |
      [ga [Ag%%]] weg |
      haal [Ag%%] $drank |
      (ga [Ag%%] naar | haal) [Ag%%] $num |
      stop;
$commando = $cmd [[Ag%%] en $cmd];
( [Ag%%] $commando [Ag%%] )
```

With such a grammar, for example the following sentences are allowed (with the English translation):

| | | |
|---|---|---|
| 1 | Achteruit en weg. | (Backward and go away.) |
| 2 | Weg. | (Go away.) |
| 3 | Stop en haal nummer drie zeven. | (Stop and fetch number 3 **7**.) |
| 4 | Naar rechts en achteruit. | (Go to the right and backward.) |
| 5 | Stop. | (Stop.) |
| 6 | Naar links en weg. | (Go to the left and go away.) |
| 7 | Haal koffie. | (Fetch coffee.) |
| 8 | Vooruit en achteruit. | (Forward and backward.) |
| 9 | Vooruit. | (Forward.) |
| 10 | Naar rechts en weg. | (Go to the right and go away.) |

In order to account for some speaker variability, a total of 10 speakers (5 male and 5 female) are included in the speech database. Each speaker spoke a total of 300 different sentences (different speaker spoke different sets of sentences), of which 250 were used for system training and 50 used for the recognition test. All the 3,000 sentence texts were generated randomly by the grammar given above. After completing the recording in quite (in a soundproof chamber), noise-corrupted sentences were simulated by adding Gaussian noise to the clean signal at 10, 16 and 22 dB SNR levels, respectively, to produce the noisy sentences. The SNR was defined as the ratio between the variances of the clean speech signal and the added noise signal, spanning the whole sentence including the silence at the beginning and the end.

The initial training used 20 sentences per speaker, all manually labelled (at word level, including marks of relatively long pauses between words). Two steps of initial training iterations were performed. For all the remaining training sentences, word-level transcriptions were obtained by manually adding long pauses between words to the sentence transcription generated by the grammar. The initial training generated HMMs with one Gaussian component per state. The observation densities were further split into 2, 4, and 8 Gaussians. After each split, one iteration of embedded training was

performed, and after the last split, 4 iterations finished the training. A total of four sets of HMMs were generated: for MFCC and RASTA-PLP, and with clean and 10 dB SNR (the most serious noise in our test), respectively

In the tests, the same finite-state grammar as above was used to constrain the between-word transitions, and no further bi-gram language model was used for such a relatively simple task. The test set consists of 500 sentences, with 50 sentences of each of the 10 speakers. Each of the four sets of HMMs was tested on four conditions: clean, 10, 16, and 22 dB SNR, respectively. This resulted in a total of 16 acoustic test conditions. Among these 16 conditions, four are "matched noise level" conditions, which tested the four sets of HMMs on the respectively same signal processing scheme and the same noise level (e.g. 10 dB MFCC). The remaining 12 "unmatched" conditions were added after finding that all the matched conditions resulted in a rather high word-correct score. We also performed another set of tests on the same 16 conditions, but without using the language model (each of the 26 words can transit to any words with the same probability). We did this to see the effect of the language model on such a simple recognition task.

The language match factors and the word-insertion penalty were optimised (producing about the same insertion and deletion errors) only to a single set of values for the 16 conditions with the language model, and to another set of values for the other 16 conditions. In the current test, environmental sounds (telephones etc.) have not yet been included.

### 3.3. Results and discussion

The recognition scores for the 16 conditions, with and without the language models, are shown in the two tables. Since the effect of using the language model (as shown in tables 1 and 2) in this simple task is not very significant, we further show in Figure 1 only the scores of the 16 conditions *with* the language model.

Table 1 & 2. Percentages (%) are shown of both the word correct score (cor.) and accuracy score (acc.), which includes the insertion errors. The shaded areas show the *matched* conditions, e.g., using the HMMs trained with clean speech also to be tested on clean speech.

| Without language model (word %) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Test SNR | Clean | | 22 dB | | 16 dB | | 10 dB | |
| HMM trained on | cor. | acc. | cor. | acc. | cor. | acc. | cor. | acc. |
| MFCC clean | 97.57 | 97.06 | 85.42 | 76.95 | 67.01 | 45.20 | 42.37 | 23.16 |
| MFCC 10 dB | 24.52 | 0.86 | 85.37 | 75.48 | 94.58 | 92.88 | 93.62 | 92.26 |
| RASTA clean | 97.51 | 97.12 | 89.44 | 71.64 | 74.41 | 50.00 | 50.11 | 35.14 |
| RASTA 10 dB | 43.79 | 28.36 | 87.63 | 82.66 | 94.01 | 92.99 | 94.01 | 93.50 |

| With language model (word %) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Test SNR | Clean | | 22 dB | | 16 dB | | 10 dB | |
| HMM trained on | cor. | acc. | cor. | acc. | cor. | acc. | cor. | acc. |
| MFCC clean | 99.83 | 99.72 | 97.57 | 96.50 | 76.95 | 66.33 | 51.36 | 36.55 |
| MFCC 10 dB | 42.09 | 19.15 | 95.76 | 92.37 | 98.70 | 98.14 | 98.47 | 98.25 |
| RASTA clean | 99.72 | 99.55 | 95.82 | 93.28 | 79.94 | 65.31 | 66.05 | 50.85 |
| RASTA 10 dB | 55.65 | 51.13 | 97.80 | 96.89 | 98.87 | 98.76 | 98.76 | 98.64 |

The following points can be observed. (1) The effect of using the language model is marginal. The following points are for the conditions with the language model, however these points apply equally well to the conditions without the language model. (2) Each of the four HMM sets performs best on its matched condition, whereas the performance degrades as the noise condition gradually deviates more from the training condition. For example, the HMM trained with clean MFCC recognises 99.83% correct of the clean speech, but it only performs 51.36% correct when tested at 10 dB SNR. (3) The scores for "matched" noisy condition (10 dB) are slightly lower than those of the "matched" clean conditions, for both MFCC and RASTA-PLP processing. (4) RASTA-PLP gives rise to a clearly higher score for the *unmatched* conditions than MFCC does. For example, for recognition of clean speech, the noise-trained (10dB) RASTA-PLP HMMs resulted in 55.65% correct whereas the noise-trained MFCC HMMs resulted in only 42.09% correct. This is a clear indication that the RASTA-PLP has a better ability to adapt to a different acoustic environment.

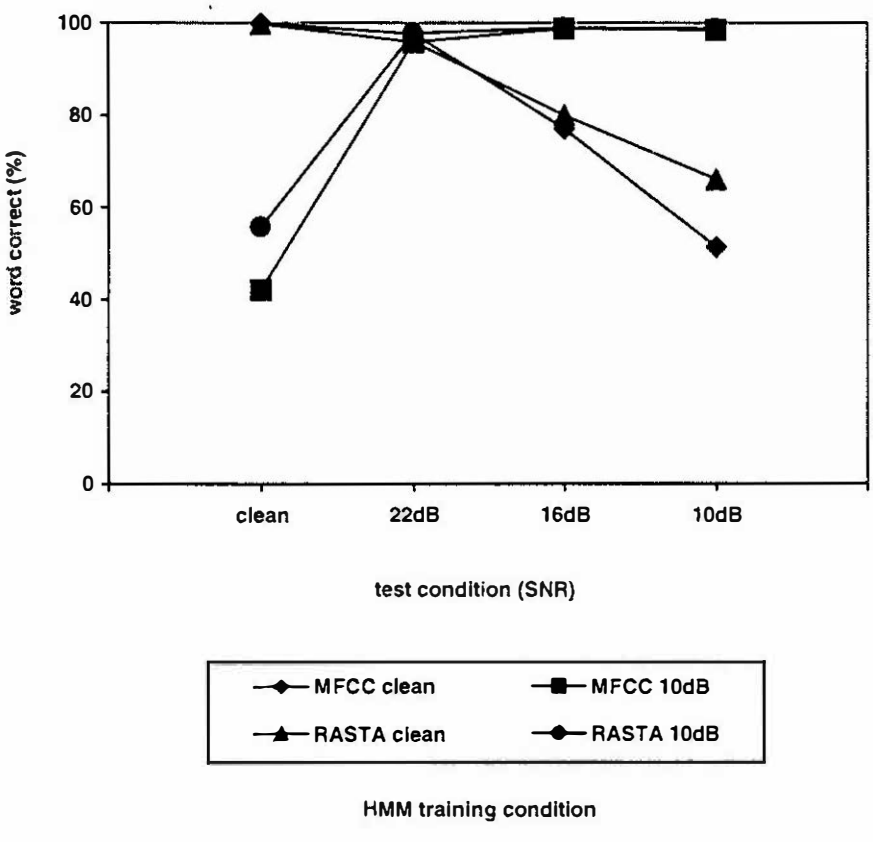


Figure 1. Word correct scores for tests using the language model.

## 4. Conclusion

In this paper we reported about a pilot study on the first investigation of robust speech recognition. The current state of the art of the technical development in ASR, particularly its application for noisy and distorted speech is reviewed. Based on this review, the research directions are pointed out for the pilot study, as well as for the speech recognition part of the long-term SUM project. The direction for the short project is proven to be a good one, as shown with the recognition results obtained from the tests. Concerning the limited tests in this study, the results should be extended only with care into the long-term SUM project. Some of the limits are, that the task is

fictitious, the environment sounds (to be recognised) are not yet included, and most of all, the environment noise is simulated and added later on, rather than from the real room acoustics. Of course, the final performance only makes sense when the whole SUM is developed and is integrated with the interactive robot(s).

The system design of the recogniser itself plays an essential role in the good performance achieved in the present tests. This design is mainly based on the previous experience of the first author (Wang, 1997), therefore those details are not presented again in this paper. In future work, more experience will be required for the more complicated, real world recognition task. A thorough understanding of the techniques in the field of ASR, as well as actively tracing new technical developments in robust recognition, will be essential for the success of the SUM project. The integration of the SUM within the robot system will depend on good and extensive corporation with other researchers in the group of intelligent autonomous systems. The feedback requirement for the recogniser from the special system configuration and acoustics of the robot will also be very valuable in providing new insight for the development of (the application of) the robust ASR systems.

## 5. Acknowledgements

The research project reported in this paper was partly supported by SION. Several useful discussions with Prof. Frans Groen and Dr. Edo Dooijes are greatly appreciated. Finally we thank all ten speakers at the IFA for their willingness of recording the 300 very boring sentences used for this study. Dr. Rob van Son is thanked for his useful comments for improving the text of this paper.

## 6. References

Gales, M.J.F. & Woodland, P.C. (1996): "Mean and variance adaptation within the MLLR framework", *Computer Speech and Language* **10**, 249-264.

Hermansky, H. (1990): "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.* **87**(4), 1738-1752.

Hermansky, H. & Morgan, N. (1994): "RASTA processing of speech", *IEEE Trans. Speech and Audio Proc.* **2**(4), 578-589.

Hermansky, H., Morgan, N., Bayya, A. & Kohn, P. (1991): "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)", *Proceedings EUROSPEECH'91,* Genova, Italy, 1367-1370.

Juang, B.-H. & Paliwal, K. K. (1992): "Hidden Markov models with first- order equalization for noisy speech recognition", *IEEE Trans. Signal Processing* **40**(9), 2136-2143.

Juang, B.H. (1995): "Recent developments in robust speech recognition", in: Ramachandran, R.P. & Mammone, R. (eds.): *Modern methods of speech processing,* Kluwer Academic Publ., Boston, 231-249.

Oh, S. & Viswanathan, V. (1995): "Microphone array for hand-free voice communication in a car", in Ramachandran, P.R. and Mammone, R. (eds): *Modern methods in speech processing,* Kluwer Acad. Publ. Boston, 351-375.

Sankar, A. & Lee, C.-H. (1996): "A maximum-likelihood approach to stochastic matching for robust speech recognition", *IEEE Trans. Speech and Audio Proc.* **4**(3), 190-202.

Stern, R.M., Acero, A., Liu, F.-H. & Ohshima, Y. (1996): "Signal processing for robust speech recognition", in Lee, C.-H., Soong, F.K. & Paliwal, K.K. (eds.): *Automatic speech and speaker recognition: advanced topics,* Kluwer Academic Publ., Boston, 357-384.

Wang, X. (1997): *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition,* Ph.D. thesis, University of Amsterdam, IFOTT series on Studies on Language and Language Use no. 29, 190 pages.

Young, S.J. (1992): *HTK: Hidden Markov model toolkit v1.4 Reference and Programming Manuals,* Cambridge Univ.