Chapter 1

Spoken Language Input

(Following section is taken from Chapter 1 "Spoken Language Input")
of the book: "Survey of the state of the art in human language technology")

1.7 Speaker Recognition

Sadaoki Furui

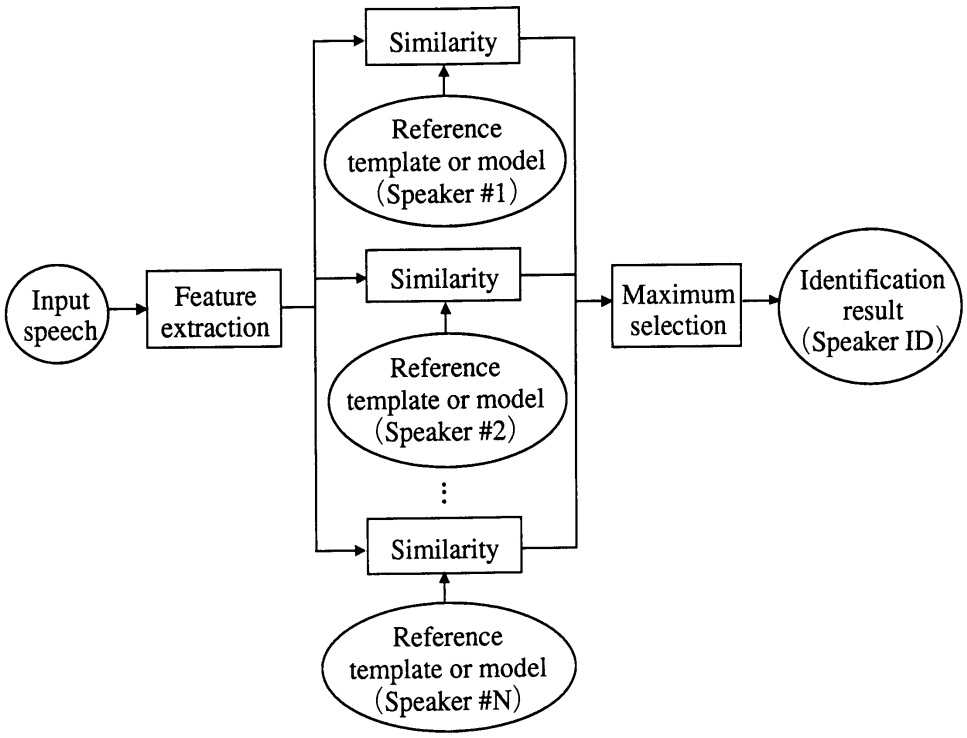
NTT Human Interface Laboratories, Tokyo, Japan

1.7.1 Principles of Speaker Recognition

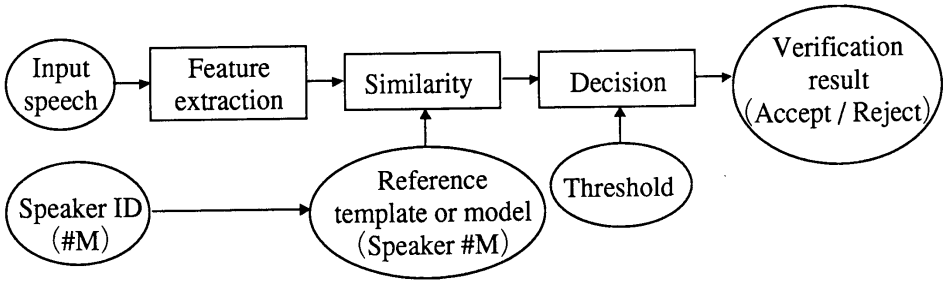
Speaker recognition, which can be classified into identification and verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. AT&T and TI (with Sprint) have started field tests and actual application of speaker recognition technology; Sprint's Voice Phone Card is already being used by many customers. In this way, speaker recognition technology is expected to create new services that will make our daily lives more convenient. Another important application of speaker recognition technology is for forensic purposes.

Figure 1.1 shows the basic structures of speaker identification and verification systems. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification.

There is also the case called *open set* identification, in which a reference model for an unknown speaker may



(a) Speaker identification



(b) Speaker verification

Figure 1.1: Basic structures of speaker recognition systems.

not exist. This is usually the case in forensic applications. In this situation, an additional decision alternative, *the unknown does not match any of the models*, is required. In both verification and identification processes, an additional threshold test can be used to determine if the match is close enough to accept the decision or if more speech data needed.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken.

Both text-dependent and independent methods share a problem however. These systems can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used. Yet even this method is not completely reliable, since it can be deceived with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted (machine-driven-text-dependent) speaker recognition method has recently been proposed by Matsui and Furui (1993b).

1.7.2 Feature Parameters

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments).

The most common short-term spectral measurements currently used are Linear Predictive Coding (LPC)-derived cepstral coefficients and their regression coefficients. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed from LPC coefficients. Therefore it provides a stabler representation from one repetition to another of a particular speaker's utterances. As for the regression coefficients, typically the first- and second-order coefficients are extracted at every frame period to represent the spectral dynamics. These coefficients are derivatives of the time functions of the cepstral coefficients and are respectively called the delta- and delta-delta-cepstral coefficients.

1.7.3 Normalization Techniques

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from trial to trial (intersession variability and variability over time). Variations arise from the speakers themselves, from differences in recording and transmission conditions, and from background noise. Speakers cannot repeat an utterance precisely the same way from trial to trial. It is well known that samples of the same utterance recorded in one session are much more highly correlated than samples recorded in separate sessions. There are also long-term changes in voices.

It is important for speaker recognition systems to accommodate these variations. Two types of normalization techniques have been tried; one in the parameter domain, the other in the distance/similarity domain.

Parameter-Domain Normalization

Spectral equalization, the so-called *blind equalization* method, is a typical normalization technique in the parameter domain that has been confirmed to be effective in reducing linear channel effects and long-term spectral variation (Atal, 1974; Furui, 1981). This method is especially effective for text-dependent speaker recognition applications that use sufficiently long utterances. Cepstral coefficients are averaged over the duration of an entire utterance and the averaged values subtracted from the cepstral coefficients of each frame. Additive variation in the log spectral domain can be compensated for fairly well by this method. However, it unavoidably removes some text-dependent and speaker specific features; therefore it is inappropriate for short utterances in speaker recognition applications.

Distance/Similarity-Domain Normalization

A normalization method for distance (similarity, likelihood) values using a likelihood ratio has been proposed by Higgins, Bahler, et al. (1991). The likelihood ratio is defined as the ratio of two conditional probabilities of the observed measurements of the utterance: the first probability is the likelihood of the acoustic data given the claimed identity of the speaker, and the second is the likelihood given that the speaker is an imposter. The likelihood ratio normalization approximates optimal scoring in the Bayes sense.

A normalization method based on *a posteriori* probability has also been proposed by Matsui and Furui (1994a). The difference between the normalization method based on the likelihood ratio and the method based on *a posteriori* probability is whether or not the claimed speaker is included in the speaker set for normalization; the speaker set used in the method based on the likelihood ratio does not include the claimed speaker, whereas the normalization term for the method based on *a posteriori* probability is calculated by using all the reference speakers, including the claimed speaker.

Experimental results indicate that the two normalization methods are almost equally effective (Matsui & Furui, 1994a). They both improve speaker separability and reduce the need for speaker-dependent or text-dependent thresholding, as compared with scoring using only a model of the claimed speaker.

A new method has recently been proposed in which the normalization term is approximated by the likelihood of a single mixture model representing the parameter distribution for all the reference speakers. An advantage of this method is that the computational cost of calculating the normalization term is very small, and this method has been confirmed to give much better results than either of the above-mentioned normalization methods (Matsui & Furui, 1994a). 1994].

1.7.4 Text-Dependent Speaker Recognition Methods

Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm, and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated.

The hidden Markov model (HMM) can efficiently model statistical variation in spectral features. Therefore, HMM-based methods were introduced as extensions of the DTW-based methods, and have achieved significantly better recognition accuracies (Naik, Netsch, et al., 1989).

1.7.5 Text-Independent Speaker Recognition Methods

One of the most successful text-independent recognition methods is based on vector quantization (VQ). In this method, VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific features. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

Temporal variation in speech signal parameters over the long term can be represented by stochastic Markovian transitions between states. Therefore, methods using an ergodic HMM, where all possible transitions between states are allowed, have been proposed. Speech segments are classified into one of the broad phonetic categories corresponding to the HMM states. After the classification, appropriate features are selected.

In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after the phonetic categorization, a comparison with the reference template for each particular category provides a verification score for that category. The final verification score is a weighted linear combination of the scores from each category.

This method was extended to the richer class of mixture autoregressive (AR) HMMs. In these models, the states are described as a linear combination (mixture) of AR sources. It can be shown that mixture models are equivalent to a larger HMM with simple states, with additional constraints on the possible transitions between states.

It has been shown that a continuous ergodic HMM method is far superior to a discrete ergodic HMM method and that a continuous ergodic HMM method is as robust as a VQ-based method when enough training data is available. However, when little data are available, the VQ-based method is more robust than a continuous HMM method (Matsui & Furui, 1993a).

A method using statistical dynamic features has recently been proposed. In this method, a multivariate autoregression (MAR) model is applied to the time series of cepstral vectors and used to characterize speakers. It was reported that identification and verification rates were almost the same as obtained by an HMM-based method (Griffin, Matsui, et al., 1994).

1.7.6 Text-Prompted Speaker Recognition Method

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested. Not only can this method accurately recognize speakers, but it can also reject utterances whose text differs from the prompted text, even if it is spoken by the registered speaker. A recorded voice can thus be correctly rejected.

This method is facilitated by using speaker-specific phoneme models as basic acoustic units. One of the major issues in applying this method is how to properly create these speaker-specific phoneme models from training utterances of a limited size. The phoneme models are represented by Gaussian-mixture continuous HMMs or tied-mixture HMMs, and they are made by adapting speaker-independent phoneme models to each speaker's voice. In order to properly adapt the models of phonemes that are not included in the training utterances, a new adaptation method based on tied-mixture HMMs was recently proposed by Matsui and Furui (1994b).

In the recognition stage, the system concatenates the phoneme models of each registered speaker to create a sentence HMM, according to the prompted text. Then, the likelihood of the input speech matching the sentence model is calculated and used for the speaker recognition decision. If the likelihood is high enough, the speaker is accepted as the claimed speaker.

1.7.7 Future Directions

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises.

From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. Studies on ways to automatically extract the speech periods of each person separately from a dialogue involving more than two people have recently appeared as an extension of speaker recognition technology.

This section was not intended to be a comprehensive review of speaker recognition technology. Rather, it was intended to give an overview of recent advances and the problems which must be solved in the future. The reader is referred to the following papers for more general reviews: Furui, 1986a; Furui, 1989; Furui, 1991; Furui, 1994; O'Shaughnessy, 1986; Rosenberg & Soong, 1991.

1.8 Chapter References

Acero, A. and Stern, R. M. (1990). Environmental robustness in automatic speech recognition. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 849–852, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Alleva, F., Huang, X., and Hwang, M. Y. (1993). An improved search algorithm using incremental knowledge for continuous speech recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 307–310, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Alvarado, V. M. and Silverman, H. F. (1990). Experimental results showing the effects of optimal spacing between elements of a linear microphone array. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 837–840, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Anastasakos, T., Makhoul, J., and Schwartz, R. (1994). Adaptation to new microphones using tied-mixture normalization. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 433–436, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Applebaum, T. H. and Hanson, B. A. (1989). Regression features for recognition of speech in quiet and in noise. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 985–988, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

ARPA (1993). *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1994). *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1995a). *Proceedings of the 1995 ARPA Human Language Technology Workshop*. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1995b). *Proceedings of the ARPA Spoken Language Systems Technology Workshop*. Advanced Research Projects Agency, Morgan Kaufmann.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312.

Aubert, X., Dugast, C., Ney, H., and Steinbiss, V. (1994). Large vocabulary continuous speech recognition of wall street journal data. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 129–132, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Bahl, L. R., Bellegarda, J. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., and Picheny, M. A. (1993). Multonic Markov word models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(3):334–344.

Bahl, L. R., Brown, P. F., de Souza, P. V., Mercer, R. L., and Picheny, M. A. (1993). A method for the construction of acoustic Markov models for words. *IEEE Transactions on Speech and Audio Processing*, 1(4):443–452.

Bahl, L. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., and Picheny, M. A. (1991). Decision trees for phonological rules in continuous speech. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 185–188, Toronto. Institute of Electrical and Electronic Engineers.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

Bellegarda, J. R., de Souza, P. V., Nadas, A. J., Nahamoo, D., Picheny, M. A., and Bahl, L. (1992). Robust speaker adaptation using a piecewise linear acoustic mapping. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 445–448, San Francisco. Institute of Electrical and Electronic Engineers.

Bengio, Y., DeMori, R., Flammia, G., and Kompe, R. (1992). Global optimization of a neural network—hidden Markov model hybrid. *IEEE Transactions on Neural Networks*, 3(2):252–259.

Berger, A., Della Pietra, S., and Della Pietra, V. (1994). Maximum entropy methods in machine translation. Technical report, IBM Research Report.

Bocchieri, E. L. (1993). Vector quantization for the efficient computation of continuous density likelihoods. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 692–694, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1990). Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, Paris, France.

Che, C., Lin, J., Pearson, J., de Vries, B., and Flanagan, J. (1994). Microphones arrays and neural networks for robust speech recognition. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Cohen, J., Gish, H., and Flanagan, J. (1994). Switchboard—the second year. Technical Report /pub/caipworks2 at ftp.rutgers.edu, CAIP Summer Workshop in Speech Recognition: Frontiers in Speech Processing II.

Cohen, J. R. (1989). Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, 85(6):2623–2629.

Cole, R. A., Hirschman, L., et al. (1992). Workshop on spoken language understanding. Technical Report CSE 92-014, Oregon Graduate Institute of Science & Technology, P.O.Box 91000, Portland, OR 97291-1000 USA.

DARPA (1990). *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1991). *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1992). *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-36(5):961–1005.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28:357–366.

DeMori, R. and Kuhn, R. (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(6):570–583.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum-likelihood from incompete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B.*, 39:1–38.

Digalakis, V. and Murveit, H. (1994). Genones: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 537–540, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Duda, R. O., Lyon, R. F., and Slaney, M. (1990). Correlograms and the separation of sounds. In *Proceedings of the 24th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 7457–7461.

Ephraim, Y. (1992). Gain-adapted hidden Markov models for recognition of clean and noisy speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 40:1303–1316.

Erell, A. and Weintraub, M. (1990). Recognition of noisy speech: Using minimum-mean log-spectral distance estimation. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 341–345, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

ESCA (1993). Proceedings of the ESCA workshop on prosody. Technical Report Working Papers 41, Lund University Department of Linguistics.

Eurospeech (1989). *Eurospeech '89, Proceedings of the First European Conference on Speech Communication and Technology*, Paris. European Speech Communication Association, European Speech Communication Association.

Eurospeech (1991). *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, Genova, Italy. European Speech Communication Association.

Eurospeech (1993). *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin. European Speech Communication Association.

Fanty, M., Barnard, E., and Cole, R. A. (1995). Alphabet recognition. In *Handbook of Neural Computation*. Publisher Unknown. In press.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272.

Furui, S. (1986a). Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, 5(2):183–197.

Furui, S. (1986b). Speaker-independent isolated word recognition using dynamic features of the speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):59–59.

Furui, S. (1989). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, New York.

Furui, S. (1991). Speaker-dependent-feature extraction, recognition and processing techniques. *Speech Communication*, 10(5-6):505–520.

Furui, S. (1994). An overview of speaker recognition technology. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 1–9.

Gales, M. J. F. and Young, S. J. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 233–236, San Francisco. Institute of Electrical and Electronic Engineers.

Gauvain, J.-L. and Lee, C.-H. (1991). Bayesian learning for hidden markov model with gaussian mixture state observation densities. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, pages 939–942, Genova, Italy. European Speech Communication Association.

Ghitza, O. (1988). Temporal non-place information in the auditory-nerve firing patterns as a front end for speech recognition in a noisy environment. *Journal of Phonetics*, 16(1):109–124.

Goddeau, D. and Zue, V. (1992). Integrating probabilistic LR parsing into speech understanding systems. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, San Francisco. Institute of Electrical and Electronic Engineers.

Greenberg, S. (1988). Theme issue: Representation of speech in the auditory periphery. *Journal of Phonetics*, 16(1).

Griffin, C., Matsui, T., and Furui, S. (1994). Distance measures for text-independent speaker recognition based on MAR model. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 309–312, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Haeb-Umbach, R., Geller, D., and Ney, H. (1993). Improvements in connected digit recognition using linear discriminant analysis and mixture densities. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 239–242, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis for speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). Compensation for the effects of the communication channel in auditory-like analysis of speech. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, pages 1367–1370, Genova, Italy. European Speech Communication Association.

Hermansky, H., Morgan, N., and Hirsch, H. G. (1993). Recognition of speech in additive and convolutional noise based on RASTA spectral processing. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 83–86, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Higgins, A. L., Bahler, L., and Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106.

Hindle, D. (1983). Deterministic parsing of syntactic nonfluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128, Cambridge, Massachusetts. Association for Computational Linguistics.

Hirsch, H. G., Meyer, P., and Ruehl, H. W. (1991). Improved speech recognition using high-pass filtering of subband envelopes. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, pages 413–416, Genova, Italy. European Speech Communication Association.

Hon, H.-W. and Lee, K.-F. (1991). CMU robust vocabulary-independent speech recognition system. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 889–892, Toronto. Institute of Electrical and Electronic Engineers.

Huang, X. D., Ariki, Y., and Jack, M. (1990). *Hidden Markov Models for Speech Recognition*. Edinburgh University Press.

Huang, X. D. and Lee, K. F. (1993). On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(2):150–157.

Hunt, M. J. (1993). Signal processing for speech. In Asher, R. E., editor, *The Encyclopedia of Language and Linguistics*. Pergamon Press.

Hunt, M. J. and Lefèbvre, C. (1989). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 262–265, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

Hwang, M. Y. and Huang, X. (1993). Shared-distribution hidden Markov models for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4):414–420.

ICASSP (1987). *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, Dallas. Institute of Electrical and Electronic Engineers.

ICASSP (1989). *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

ICASSP (1990). *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

ICASSP (1991). *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, Toronto. Institute of Electrical and Electronic Engineers.

ICASSP (1992). *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, San Francisco. Institute of Electrical and Electronic Engineers.

ICASSP (1993). *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

ICASSP (1994). *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

ICSLP (1990). *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan.

ICSLP (1992). *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, Alberta, Canada. University of Alberta.

ICSLP (1994). *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, Japan.

Iyer, R., Ostendorf, M., and Rohlicek, R. (1994). An improved language model using a mixture of Markov components. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Jelinek, F. (1969). A fast sequential decoding algorithm using a stack. *IBM journal of Research and Development*, 13.

Jelinek, F., Merialdo, B., Roukos, S., and Strauss, M. (1991). A dynamic language model for speech recognition. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 293–295, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Juang, B. H. (1991). Speech recognition in adverse environments. *Computer Speech and Language*, pages 275–294.

Juang, B. H., Rabiner, L. R., and Wilpon, J. G. (1986). On the use of bandpass liftering in speech recognition. In *Proceedings of the 1986 International Conference on Acoustics, Speech, and Signal Processing*, pages 765–768, Tokyo. Institute of Electrical and Electronic Engineers.

Koehler, J., Morgan, N., Hermansky, H., Hirsch, H. G., and Tong, G. (1994). Integrating RASTA-PLP into speech recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 421–424, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., and Zavaliagkos, G. (1994). Comparative experiments on large vocabulary speech recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 561–564, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Kuhn, R., De Mori, R., and Millien, E. (1994). Learning consistent semantics from training data. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 37–40, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Lafferty, J., Sleator, D., and Temperley, D. (1992). Grammatical trigrams: A probabilistic model of link grammar. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*.

Lau, R., Rosenfeld, R., and Roukos, S. (1993). Trigger-based language models: A maximum entropy approach. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 45–48, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Lickley, R. J. (1994). *Detecting Disfluency in Spontaneous Speech*. PhD thesis, University of Edinburgh, Scotland.

Lim, J. and Oppenheim, A. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67:1586–1604.

Lippmann, R. P., Martin, F. A., and Paul, D. B. (1987). Multi-style training for robust isolated-word speech recognition. In *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, pages 709–712, Dallas. Institute of Electrical and Electronic Engineers.

Liu, F.-H., Stern, R. M., Acero, A., and Moreno, P. (1994). Environment normalization for robust speech recognition using direct cepstral comparison. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 61–64, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Lockwood, P., Boudy, J., and Blanchet, M. (1992). Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 265–268, San Francisco. Institute of Electrical and Electronic Engineers.

Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. In *Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing*, pages 1282–1285. Institute of Electrical and Electronic Engineers.

Markel, J. D. and Gray, Jr., A. H. (1976). *Linear Prediction of Speech*. Springer-Verlag, Berlin.

Matsui, T. and Furui, S. (1993a). Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 157–160, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Matsui, T. and Furui, S. (1993b). Concatenated phoneme models for text-variable speaker recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 391–394, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Matsui, T. and Furui, S. (1994a). Similarity normalization method for speaker verification based on a posteriori probability. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 59–62.

Matsui, T. and Furui, S. (1994b). Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 125–128, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Meng, H. M. and Zue, V. W. (1990). A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, volume 2, pages 1053–1056, Kobe, Japan.

Merhav, N. and Ephraim, Y. (1991). Maximum likelihood hidden markov modeling using a dominant state sequence of states. *IEEE Transactions on Signal Processing*, 39(9):2111–2114.

Murveit, H., Butzberger, J., Digilakis, V., and Weintraub, M. (1993). Large-vocabulary dictation using SRI's DECIPHER speech recognition system: Progressive search techniques. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 319–322, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Naik, J. M., Netsch, L. P., and Doddington, G. R. (1989). Speaker verification over long distance telephone lines. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 524–527, Glasgow, Scotland. Institute of Electrical and Electronic Engineers.

Neumeyer, L. and Weintraub, M. (1994). Probabilistic optimum filtering for robust speech recognition. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 417–420, Adelaide, Australia. Institute of Electrical and Electronic Engineers.

Ney, H., Mergel, D., Noll, A., and Paesler, A. (1992). Data driven search organization for continuous speech recognition. *IEEE Transactions on Signal Processing*, 40(2):272–281.

Nilsson, N. J. (1971). *Problem-Solving Methods in Artificial Intelligence*. McGraw-Hill, New York.

Ohshima, Y. (1993). *Robustness in Speech Recognition using Physiologically-Motivated Signal Processing*. PhD thesis, CMU.

O'Shaughnessy, D. (1986). Speaker recognition. *IEEE Acoustics, Speech and Signal Processing Magazine*, 3(4):4–17.

Pallett, D. (1991). DARPA resource management and ATIS benchmark test poster session. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 49–58, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D. (1992). ATIS benchmarks. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Dahlgren, N., Fiscus, J., Fisher, W., Garofolo, J., and Tjaden, B. (1992). DARPA February 1992 ATIS benchmark test results. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*, pages 15–27. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Fiscus, J., Fisher, W., and Garofolo, J. (1993). Benchmark tests for the DARPA spoken language program. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, pages 7–18, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., and Prysbocki, M. (1994). 1993 benchmark tests for the ARPA spoken language program. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages 49–74, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D., Fisher, W., Fiscus, J., and Garofolo, J. (1990). DARPA ATIS test results. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 114–121, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Pallett, D. S., Fiscus, J. G., Fisher, W. M., Garofolo, J. S., Lund, B. A., Martin, A., and Przybocki, M. A. (1995). 1994 benchmark tests for the ARPA spoken language program. In *Proceedings of the 1995 ARPA Human Language Technology Workshop*, pages 5–36. Advanced Research Projects Agency, Morgan Kaufmann.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1991). Complex sounds and auditory images. In *Auditory Physiology and Perception*, pages 429–446. Pergamon Press.

Paul, D. B. (1994). The Lincoln large-vocabulary stack-decoder based HMM CSR. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages 374–379, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Peterson, P. M. (1989). Adaptive array processing for multiple microphone hearing aids. Technical Report 541, Research Laboratory of Electronics, MIT, Cambridge, Massachusetts.

Porter, J. E. and Boll, S. F. (1984). Optimal estimators for spectral restoration of noisy speech. In *Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing*, pages 18.A.2.1–4. Institute of Electrical and Electronic Engineers.

Price, P. and Ostendorf, M. (1995). Combining linguistic with statistical methods in modeling prosody. In Morgan, J. L. and Demuth, K., editors, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Signal Processing. Prentice-Hall, Englewood Cliffs, New Jersey.

Rosenberg, A. E. and Soong, F. K. (1991). Recent research in automatic speaker recognition. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, pages 701–737. Marcel Dekker, New York.

Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150.

Schwartz, R., Chow, Y., and Kubala, F. (1987). Rapid speaker adaption using a probabalistic spectral mapping. In *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, pages 633–636, Dallas. Institute of Electrical and Electronic Engineers.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):55–76.

Shirai, K. and Furui, S. (1994). Special issue on spoken dialogue. *Speech Communication*, 15(3-4).

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, Stanford University.

Shukat-Talamazzini, E. G., Niemann, H., Eckert, W., Kuhn, T., and Rieck, S. (1992). Acoustic modeling of subword units in the ISADORA speech recognizer. In *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 577–580, San Francisco. Institute of Electrical and Electronic Engineers.

Stockham, T. G., J., Connon, T. M., and Ingebretsen, R. B. (1975). Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692.

Sullivan, T. M. and Stern, R. M. (1993). Multi-microphone correlation-based processing for robust speech recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 91–94, Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.

Van Compernolle, D. (1990). Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 833–836, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Varga, A. P. and Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 845–848, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Waibel, A. and Lee, K. F. (1990). *Readings in Speech Recognition*. Morgan Kaufmann.

Zue, V., Glass, J., Phillips, M., and Seneff, S. (1990). The MIT SUMMIT speech recognition system: A progress report. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.