The Application of Hidden Markov Models in Speech Recognition

The Application of Hidden Markov Models in Speech Recognition

Mark Gales

Cambridge University Engineering Department
Cambridge
CB2 1PZ
UK
mjfg@eng.cam.ac.uk

Steve Young

Cambridge University Engineering Department
Cambridge
CB2 1PZ
UK
sjy@eng.cam.ac.uk

the essence of knowledge

Boston – Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is M. Gales and S. Young, The Application of Hidden Markov Models in Speech Recognition, Foundations and Trends® in Signal Processing, vol 1, no 3, pp 195–304, 2007

ISBN: 978-1-60198-120-2

Foundations and Trends® in Signal Processing
Volume 1 Issue 3, 2007

Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
- Classification and detection
- Estimation and regression
- Tree-structured methods

Information for Librarians

Foundations and Trends® in Signal Processing, 2007, Volume 1, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in
Signal Processing
Vol. 1, No. 3 (2007) 195–304

DOI: 10.1561/2000000004

The Application of Hidden Markov Models in Speech Recognition

Mark Gales[1] and Steve Young[2]

[1] *Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK, mjfg@eng.cam.ac.uk*
[2] *Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK, sjy@eng.cam.ac.uk*

Abstract

Hidden Markov Models (HMMs) provide a simple and effective framework for modelling time-varying spectral vector sequences. As a consequence, almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMMs.

Whereas the basic principles underlying HMM-based LVCSR are rather straightforward, the approximations and simplifying assumptions involved in a direct implementation of these principles would result in a system which has poor accuracy and unacceptable sensitivity to changes in operating environment. Thus, the practical application of HMMs in modern systems involves considerable sophistication.

The aim of this review is first to present the core architecture of a HMM-based LVCSR system and then describe the various refinements which are needed to achieve state-of-the-art performance. These

refinements include feature projection, improved covariance modelling, discriminative parameter estimation, adaptation and normalisation, noise compensation and multi-pass system combination. The review concludes with a case study of LVCSR for Broadcast News and Conversation transcription in order to illustrate the techniques described.

Contents

1

Introduction

Automatic continuous speech recognition (CSR) has many potential applications including command and control, dictation, transcription of recorded speech, searching audio documents and interactive spoken dialogues. The core of all speech recognition systems consists of a set of statistical models representing the various sounds of the language to be recognised. Since speech has temporal structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range, the hidden Markov model (HMM) provides a natural framework for constructing such models [13].

HMMs lie at the heart of virtually all modern speech recognition systems and although the basic framework has not changed significantly in the last decade or more, the detailed modelling techniques developed within this framework have evolved to a state of considerable sophistication (e.g. [40, 117, 163]). The result has been steady and significant progress and it is the aim of this review to describe the main techniques by which this has been achieved.

The foundations of modern HMM-based continuous speech recognition technology were laid down in the 1970's by groups at Carnegie-Mellon and IBM who introduced the use of discrete density HMMs

[11, 77, 108], and then later at Bell Labs [80, 81, 99] where continuous density HMMs were introduced.[1] An excellent tutorial covering the basic HMM technologies developed in this period is given in [141].

Reflecting the computational power of the time, initial development in the 1980's focussed on either discrete word speaker dependent large vocabulary systems (e.g. [78]) or whole word small vocabulary speaker independent applications (e.g. [142]). In the early 90's, attention switched to continuous speaker-independent recognition. Starting with the artificial 1000 word *Resource Management* task [140], the technology developed rapidly and by the mid-1990's, reasonable accuracy was being achieved for unrestricted speaker independent dictation. Much of this development was driven by a series of DARPA and NSA programmes [188] which set ever more challenging tasks culminating most recently in systems for multilingual transcription of broadcast news programmes [134] and for spontaneous telephone conversations [62].

Many research groups have contributed to this progress, and each will typically have its own architectural perspective. For the sake of logical coherence, the presentation given here is somewhat biassed towards the architecture developed at Cambridge University and supported by the HTK Software Toolkit [189].[2]

The review is organised as follows. Firstly, in *Architecture of a HMM-Based Recogniser* the key architectural ideas of a typical HMM-based recogniser are described. The intention here is to present an overall system design using very basic acoustic models. In particular, simple single Gaussian diagonal covariance HMMs are assumed. The following section *HMM Structure Refinements* then describes the various ways in which the limitations of these basic HMMs can be overcome, for example by transforming features and using more complex HMM output distributions. A key benefit of the statistical approach to speech recognition is that the required models are trained automatically on data.

[1] This very brief historical perspective is far from complete and out of necessity omits many other important contributions to the early years of HMM-based speech recognition.

[2] Available for free download at `htk.eng.cam.ac.uk`. This includes a recipe for building a state-of-the-art recogniser for the Resource Management task which illustrates a number of the approaches described in this review.

The section *Parameter Estimation* discusses the different objective functions that can be optimised in training and their effects on performance. Any system designed to work reliably in real-world applications must be robust to changes in speaker and the environment. The section on *Adaptation and Normalisation* presents a variety of generic techniques for achieving robustness. The following section *Noise Robustness* then discusses more specialised techniques for specifically handling additive and convolutional noise. The section *Multi-Pass Recognition Architectures* returns to the topic of the overall system architecture and explains how multiple passes over the speech signal using different model combinations can be exploited to further improve performance. This final section also describes some actual systems built for transcribing English, Mandarin and Arabic in order to illustrate the various techniques discussed in the review. The review concludes in *Conclusions* with some general observations and conclusions.

References

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1993.

[2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proceedings of ICSLP*, Beijing, China, 2000.

[3] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic broadcast news transcription at BBN," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005.

[4] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 11, no. 3, pp. 187–206, 1997.

[5] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proceedings of ICSLP*, Philadelphia, 1996.

[6] J. A. Arrowood, *Using Observation Uncertainty for Robust Speech Recognition*. PhD thesis, Georgia Institute of Technology, 2003.

[7] X. Aubert and H. Ney, "Large vocabulary continuous speech recognition using word graphs," in *Proceedings of ICASSP*, vol. 1, pp. 49–52, Detroit, 1995.

[8] S. Axelrod, R. Gopinath, and P. Olsen, "Modelling with a subspace constraint on inverse covariance matrices," in *Proceedings of ICSLP*, Denver, CO, 2002.

[9] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings of ICASSP*, pp. 49–52, Tokyo, 1986.

[10] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Robust methods for using context-dependent features and models

in a continuous speech recognizer," in *Proceedings of ICASSP*, pp. 533–536, Adelaide, 1994.

[11] J. K. Baker, "The Dragon system — An overview," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

[12] J. Barker, M. Cooke, and P. D. Green, "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise," in *Proceedings of Eurospeech*, pp. 213–216, Aalberg, Denmark, 2001.

[13] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of American Mathematical Society*, vol. 73, pp. 360–363, 1967.

[14] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[15] P. Beyerlein, "Discriminative model combination," in *Proceedings of ASRU*, Santa Barbara, 1997.

[16] J. A. Bilmes, "Buried Markov models: A graphical-modelling approach to automatic speech recognition," *Computer Speech and Language*, vol. 17, no. 2–3, 2003.

[17] J. A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language: Processing Institute of Mathematical Analysis Volumes in Mathematics Series,* Springer-Verlag, 2003.

[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 27, pp. 113–120, 1979.

[19] C. Breslin and M. J. F. Gales, "Complementary system generation using directed decision trees," in *Proceedings of ICSLP*, Toulouse, 2006.

[20] P. F. Brown, *The Acoustic-Modelling Problem in Automatic Speech Recognition*. PhD thesis, Carnegie Mellon, 1987.

[21] P. F. Brown, V. J. Della Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer, "Class-based N-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[22] W. Byrne, "Minimum Bayes risk estimation and decoding in large vocabulary continuous speech recognition," *IEICE Transactions on Information and Systems: Special Issue on Statistical Modelling for Speech Recognition*, vol. E89-D(3), pp. 900–907, 2006.

[23] H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proceedings of ICASSP*, Montreal, Canada, March 2004.

[24] K. T. Chen, W. W. Liao, H. M. Wang, and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proceedings of ICSLP*, Beijing, China, 2000.

[25] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modelling," *Computer Speech and Language*, vol. 13, pp. 359–394, 1999.

[26] S. S. Chen and R. Gopinath, "Gaussianization," in *NIPS 2000*, Denver, CO, 2000.

[27] S. S. Chen and R. A. Gopinath, "Model selection in acoustic modelling," in *Proceedings of Eurospeech*, pp. 1087–1090, Rhodes, Greece, 1997.

[28] W. Chou, "Maximum *a-posterior* linear regression with elliptical symmetric matrix variants," in *Proceedings of ICASSP*, pp. 1–4, Phoenix, USA, 1999.

[29] W. Chou, C. H. Lee, and B. H. Juang, "Minimum error rate training based on N-best string models," in *Proceedings of ICASSP*, pp. 652–655, Minneapolis, 1993.

[30] M. Cooke, P. D. Green, and M. D. Crawford, "Handling missing data in speech recognition," in *Proceedings of ICSLP*, pp. 1555–1558, Yokohama, Japan, 1994.

[31] M. Cooke, A. Morris, and P. D. Green, "Missing data techniques for robust speech recognition," in *Proceedings of ICASSP*, pp. 863–866, Munich, Germany, 1997.

[32] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, pp. 1–38, 1977.

[34] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proceedings of ICSLP*, pp. 806–809, Beijing, China, 2000.

[35] V. Diakoloukas and V. Digalakis, "Maximum likelihood stochastic transformation adaptation of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 177–187, 1999.

[36] V. Digalakis, H. Collier, S. Berkowitz, A. Corduneanu, E. Bocchieri, A. Kannan, C. Boulis, S. Khudanpur, W. Byrne, and A. Sankar, "Rapid speech recognizer adaptation to new speakers," in *Proceedings of ICASSP*, pp. 765–768, Phoenix, USA, 1999.

[37] C. Dimitrakakis and S. Bengio, "Boosting HMMs with an application to speech recognition," in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[38] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proceedings of ICASSP*, Orlando, FL, 2002.

[39] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the aurora 2 database," in *Proceedings of Eurospeech*, pp. 217–220, Aalberg, Denmark, 2001.

[40] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[41] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proceedings of ICASSP*, pp. 1655–1658, Istanbul, Turkey, 2000.

[42] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proceedings of Speech Transcription Workshop*, Baltimore, 2000.

[43] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," in *Proceedings of IEEE ASRU Workshop*, pp. 347–352, Santa Barbara, 1997.

[44] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996.

[45] Y. Freund and R. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, pp. 119–139, 1997.

[46] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[47] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions ASSP*, vol. 34, pp. 52–59, 1986.

[48] M. J. F. Gales, *Model-based Techniques for Noise Robust Speech Recognition*. PhD thesis, Cambridge University, 1995.

[49] M. J. F. Gales, "Transformation smoothing for speaker and environmental adaptation," in *Proceedings of Eurospeech*, pp. 2067–2070, Rhodes, Greece, 1997.

[50] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[51] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[52] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 2000.

[53] M. J. F. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 37–47, 2002.

[54] M. J. F. Gales, "Discriminative models for speech recognition," in *ITA Workshop*, University San Diego, USA, February 2007.

[55] M. J. F. Gales, B. Jia, X. Liu, K. C. Sim, P. Woodland, and K. Yu, "Development of the CUHTK 2004 RT04 Mandarin conversational telephone speech transcription system," in *Proceedings of ICASSP*, Philadelphia, PA, 2005.

[56] M. J. F. Gales, D. Y. Kim, P. C. Woodland, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, September 2006.

[57] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, vol. 12, no. 3, pp. 231–239, 1993.

[58] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, no. 4, pp. 289–308, 1995.

[59] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[60] Y. Gao, M. Padmanabhan, and M. Picheny, "Speaker adaptation based on pre-clustering training speakers," in *Proceedings of EuroSpeech*, pp. 2091–2094, Rhodes, Greece, 1997.

[61] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation of multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[62] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD," in *Proceedings of ICASSP*, vol. 1, pp. 517–520, 1992. San Francisco.

[63] V. Goel, S. Kumar, and B. Byrne, "Segmental minimum Bayes-risk ASR voting strategies," in *Proceedings of ICSLP*, Beijing, China, 2000.

[64] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "A generalisation of the Baum algorithm to rational objective functions," in *Proceedings of ICASSP*, vol. 12, pp. 631–634, Glasgow, 1989.

[65] R. Gopinath, "Maximum likelihood modelling with Gaussian distributions for classification," in *Proceedings of ICASSP*, pp. II–661–II–664, Seattle, 1998.

[66] R. A. Gopinath, M. J. F. Gales, P. S. Gopalakrishnan, S. Balakrishnan-Aiyer, and M. A. Picheny, "Robust speech recognition in noise - performance of the IBM continuous speech recognizer on the ARPA noise spoke task," in *Proceedings of ARPA Workshop on Spoken Language System Technology*, Austin, TX, 1999.

[67] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005.

[68] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, pp. 13–16, Tokyo, 1992.

[69] R. Haeb-Umbach and H. Ney, "Improvements in time-synchronous beam search for 10000-word continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 353–356, 1994.

[70] T. Hain, "Implicit pronunciation modelling in ASR," in *ISCA ITRW PMLA*, 2002.

[71] T. Hain, P. C. Woodland, T. R. Niesler, and E. W. D. Whittaker, "The 1998 HTK System for transcription of conversational telephone speech," in *Proceedings of ICASSP*, pp. 57–60, Phoenix, 1999.

[72] D. Hakkani-Tur, F. Bechet, G. Riccardi, and G. Tur, "Beyond ASR 1-best: Using word confusion networks in spoken language understanding ," *Computer Speech and Language*, vol. 20, no. 4, October 2006.

[73] T. J. Hazen and J. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," in *Proceedings of Eurospeech*, pp. 2047–2050, 1997.

[74] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[75] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, 1994.

[76] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM Journal on Research and Development*, vol. 13, 1969.

[77] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[78] F. Jelinek, "A discrete utterance recogniser," *Proceedings of IEEE*, vol. 73, no. 11, pp. 1616–1624, 1985.

[79] H. Jiang, X. Li, and X. Liu, "Large margin hidden Markov models for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1584–1595, September 2006.

[80] B.-H. Juang, "On the hidden Markov model and dynamic time warping for speech recognition — A unified view," *AT and T Technical Journal*, vol. 63, no. 7, pp. 1213–1243, 1984.

[81] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT and T Technical Journal*, vol. 64, no. 6, pp. 1235–1249, 1985.

[82] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 14, no. 4, 1992.

[83] B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 307–309, 1986.

[84] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proceedings of ICSLP*, Beijing, China, 2000.

[85] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recogniser," *IEEE Transactions on ASSP*, vol. 35, no. 3, pp. 400–401, 1987.

[86] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proceedings of EuroSpeech*, pp. 2725–2728, September 1999.

[87] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. E. Tranter, L. Wang, and P. C. Woodland, "Recent advances in broadcast news transcription," in *Proceedings of IEEE ASRU Workshop*, (St. Thomas, U.S. Virgin Islands), pp. 105–110, November 2003.

[88] D. Y. Kim, S. Umesh, M. J. F. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proceedings of ICSLP*, Jeju, Korea, 2004.

[89] N. G. Kingsbury and P. J. W. Rayner, "Digital filtering using logarithmic arithmetic," *Electronics Letters*, vol. 7, no. 2, pp. 56–58, 1971.

[90] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modelling," in *Proceedings of Eurospeech*, pp. 973–976, Berlin, 1993.

[91] R. Kuhn, L. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Finke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proceedings of ICSLP*, Sydney, 1998.

[92] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[93] H.-K. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, no. 3, pp. 873–881, 2006.

[94] L. Lamel and J.-L. Gauvain, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[95] M. I. Layton and M. J. F. Gales, "Augmented statistical models for speech recognition," in *Proceedings of ICASSP*, Toulouse, 2006.

[96] L. Lee and R. C. Rose, "Speaker normalisation using efficient frequency warping procedures," in *Proceedings of ICASSP*, Atlanta, 1996.

[97] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[98] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, pp. 29–45, 1986.

[99] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Systems Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[100] J. Li, M. Siniscalchi, and C.-H. Lee, "Approximate test risk minimization through soft margin training," in *Proceedings of ICASSP*, Honolulu, USA, 2007.

[101] H. Liao, *Uncertainty Decoding For Noise Robust Speech Recognition*. PhD thesis, Cambridge University, 2007.

[102] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proceedings of Interspeech*, pp. 3129–3132, Lisbon, Portugal, 2005.

[103] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust speech recognition," in *Proceedings of ICSLP*, Pittsburgh, PA, 2006.

[104] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noise data," in *Proceedings of ICASSP*, Honolulu, USA, 2007.

[105] R. P. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," in *Proceedings of Eurospeech*, pp. KN37–KN40, Rhodes, Greece, 1997.

[106] X. Liu and M. J. F. Gales, "Automatic model complexity control using marginalized discriminative growth functions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1414–1424, May 2007.

[107] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2, pp. 215–228, 1992.

[108] B. T. Lowerre, *The Harpy Speech Recognition System*. PhD thesis, Carnegie Mellon, 1976.

[109] X. Luo and F. Jelinek, "Probabilistic classification of HMM states for large vocabulary," in *Proceedings of ICASSP*, pp. 2044–2047, Phoenix, USA, 1999.

[110] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amount of broadcast news data," in *Proceedings of ICASSP*, pp. 1056–1059, Toulouse, May 2006.

[111] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005.

[112] B. Mak and R. Hsiao, "Kernel eigenspace-based MLLR adaptation," *IEEE Transactions Speech and Audio Processing*, March 2007.

[113] B. Mak, J. T. Kwok, and S. Ho, "Kernel eigenvoices speaker adaptation," *IEEE Transactions Speech and Audio Processing*, vol. 13, no. 5, pp. 984–992, September 2005.

[114] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimisation," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[115] S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," in *Proceedings of Eurospeech*, vol. 2, pp. 1253–1256, Madrid, 1995.

[116] A. Matsoukas, T. Colthurst, F. Richardson, O. Kimball, C. Quillen, A. Solomonoff, and H. Gish, "The BBN 2001 English conversational speech system," in *Presentation at 2001 NIST Large Vocabulary Conversational Speech Workshop*, 2001.

[117] S. Matsoukas, J.-L. Gauvain, A. Adda, T. Colthurst, C. I. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Z. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and B. Xiang, "Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1541–1556, September 2006.

[118] T. Matsui and S. Furui, "N-best unsupervised speaker adaptation for speech recognition," *Computer Speech and Language*, vol. 12, pp. 41–50, 1998.

[119] E. McDermott, T. J. Hazen, J. Le Roux, A. Nakamura, and K. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 203–223, 2007.

[120] J. McDonough, W. Byrne, and X. Luo, "Speaker normalisation with all pass transforms," in *Proceedings of ICSLP*, Sydney, 1998.

[121] C. Meyer, "Utterance-level boosting of HMM speech recognisers," in *Proceedings of ICASSP*, Orlando, FL, 2002.

[122] M. Mohri, F. Pereira, and M. Riley, "Weighted finite state transducers in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 69–88, 2002.

[123] P. J. Moreno, *Speech recognition in noisy environments*. PhD thesis, Carnegie Mellon University, 1996.

[124] P. J. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of ICASSP*, pp. 733–736, Atlanta, 1996.

[125] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional

maximum likelihood," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 31, no. 4, pp. 814–817, 1983.

[126] L. R. Neumeyer, A. Sankar, and V. V. Digalakis, "A comparative study of speaker adaptation techniques," in *Proceedings of Eurospeech*, pp. 1127–1130, Madrid, 1995.

[127] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech and Language*, vol. 8, no. 1, pp. 1–38, 1994.

[128] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," in *Proceedings of ICASSP*, Montreal, Canada, March 2004.

[129] Y. Normandin, "An improved MMIE training algorithm for speaker independent, small vocabulary, continuous speech recognition," in *Proceedings of ICASSP*, Toronto, 1991.

[130] J. J. Odell, V. Valtchev, P. C. Woodland, and S. J. Young, "A one-pass decoder design for large vocabulary recognition," in *Proceedings of Human Language Technology Workshop*, pp. 405–410, Plainsboro NJ, Morgan Kaufman Publishers Inc., 1994.

[131] P. Olsen and R. Gopinath, "Modelling inverse covariance matrices by basis expansion," in *Proceedings of ICSLP*, Denver, CO, 2002.

[132] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, 1997.

[133] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," in *Proceedings of ITRW ASR2000: ASR Challenges for the New Millenium*, pp. 128–132, Paris, 2000.

[134] D. S. Pallet, J. G. Fiscus, J. Garofolo, A. Martin, and M. Przybocki, "1998 broadcast news benchmark test results: English and non-English word error rate performance measures," Tech. Rep., National Institute of Standards and Technology (NIST), 1998.

[135] D. B. Paul, "Algorithms for an optimal A* search and linearizing the search in the stack decoder," in *Proceedings of ICASSP*, pp. 693–996, Toronto, 1991.

[136] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, 2004.

[137] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Proceedings of EuroSpeech*, Geneva, Switzerland, September 2003.

[138] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proceedings of ICASSP*, Philadelphia, 2005.

[139] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of ICASSP*, Orlando, FL, 2002.

[140] P. J. Price, W. Fisher, J. Bernstein, and D. S. Pallet, "The DARPA 1000-word Resource Management database for continuous speech recognition," *Proceedings of ICASSP*, vol. 1, pp. 651–654, New York, 1988.

[141] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[142] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using HMMs with continuous mixture densities," *AT and T Technical Journal*, vol. 64, no. 6, pp. 1211–1233, 1985.

[143] B. Raj and R. Stern, "Missing feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[144] F. Richardson, M. Ostendorf, and J. R. Rohlicek, "Lattice-based search strategies for large vocabulary recognition," in *Proceedings of ICASSP*, vol. 1, pp. 576–579, Detroit, 1995.

[145] B. Roark, M. Saraclar, and M. Collins, "Discriminative N-gram language modelling," *Computer Speech and Language*, vol. 21, no. 2, 2007.

[146] A. V. Rosti and M. Gales, "Factor analysed hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 18, no. 2, pp. 181–200, 2004.

[147] M. J. Russell and R. K. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Proceedings of ICASSP*, pp. 5–8, Tampa, FL, 1985.

[148] G. Saon, A. Dharanipragada, and D. Povey, "Feature space Gaussianization," in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[149] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proceedings of ICASSP*, Istanbul, 2000.

[150] G. Saon, D. Povey, and G. Zweig, "CTS decoding improvements at IBM," in *EARS STT workshop*, St. Thomas, U.S. Virgin Islands, December 2003.

[151] L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 115–125, 2000.

[152] R. Schluter, B. Muller, F. Wessel, and H. Ney, "Interdependence of language models and discriminative training," in *Proceedings of IEEE ASRU Workshop*, pp. 119–122, Keystone, CO, 1999.

[153] R. Schwartz and Y.-L. Chow, "A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses," in *Proceedings of ICASSP*, pp. 701–704, Toronto, 1991.

[154] H. Schwenk, "Using boosting to improve a hybrid HMM/Neural-Network speech recogniser," in *Proceedings of ICASSP*, Phoenix, 1999.

[155] M. Seltzer, B. Raj, and R. Stern, "A Bayesian framework for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[156] F. Sha and L. K. Saul, "Large margin Gaussian mixture modelling for automatic speech recognition," in *Advances in Neural Information Processing Systems*, pp. 1249–1256, 2007.

[157] K. Shinoda and C. H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proceedings of ASRU'97*, Santa Barbara, 1997.

[158] K. C. Sim and M. J. F. Gales, "Basis superposition precision matrix models for large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[159] K. C. Sim and M. J. F. Gales, "Discriminative semi-parametric trajectory models for speech recognition," *Computer Speech and Language*, vol. 21, pp. 669–687, 2007.

[160] R. Sinha, M. J. F. Gales, D. Y. Kim, X. Liu, K. C. Sim, and P. C. Woodland, "The CU-HTK Mandarin broadcast news transcription system," in *Proceedings of ICASSP*, Toulouse, 2006.

[161] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Proceedings of InterSpeech*, Lisbon, Portugal, September 2005.

[162] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *Proceedings of ICASSP*, Philadelphia, 2005.

[163] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proceedings of ICASSP*, Philadelphia, PA, 2005.

[164] S. Srinivasan and D. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Transactions of Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2130–2140, 2007.

[165] A. Stolcke, E. Brill, and M. Weintraub, "Explicit word error minimization in N-Best list rescoring," in *Proceedings of EuroSpeech*, Rhodes, Greece, 1997.

[166] B. Tasker, *Learning Structured Prediction Models: A Large Margin Approach*. PhD thesis, Stanford University, 2004.

[167] H. Thompson, "Best-first enumeration of paths through a lattice — An active chart parsing solution," *Computer Speech and Language*, vol. 4, no. 3, pp. 263–274, 1990.

[168] K. Tokuda, H. Zen, and A. W. Black, *Text to speech synthesis: New paradigms and advances*. Chapter HMM-Based Approach to Multilingual Speech Synthesis, Prentice Hall, 2004.

[169] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory model," in *Proceedings Beyond HMM Workshop*, Tokyo, December 2004.

[170] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarisation systems," *IEEE Transactions Speech and Audio Processing*, vol. 14, no. 5, September 2006.

[171] S. Tsakalidis, V. Doumpiotis, and W. J. Byrne, "Discriminative linear transforms for feature normalisation and speaker adaptation in HMM estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 367–376, 2005.

[172] L. F. Uebel and P. C. Woodland, "Improvements in linear transform based speaker adaptation," in *Proceedings of ICASSP*, pp. 49–52, Seattle, 2001.

[173] V. Valtchev, J. Odell, P. C. Woodland, and S. J. Young, "A novel decoder design for large vocabulary recognition," in *Proceedings of ICSLP*, Yokohama, Japan, 1994.

[174] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, pp. 303–314, 1997.

[175] V. Vanhoucke and A. Sankar, "Mixtures of inverse covariances," in *Proceedings of ICASSP*, Montreal, Canada, 2003.

[176] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proceedings of ICASSP*, pp. 845–848, Albuqerque, 1990.

[177] A. J. Viterbi, "Error bounds for convolutional codes and asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, 1982.

[178] F. Wallhof, D. Willett, and G. Rigoll, "Frame-discriminative and confidence-driven adaptation for LVCSR," in *Proceedings of ICASSP*, pp. 1835–1838, Istanbul, 2000.

[179] L. Wang, M. J. F. Gales, and P. C. Woodland, "Unsupervised training for Mandarin broadcast news and conversation transcription," in *Proceedings of ICASSP*, Honolulu, USA, 2007.

[180] L. Wang and P. Woodland, "Discriminative adaptive training using the MPE criterion," in *Proceedings of ASRU*, St Thomas, U.S. Virgin Islands, 2003.

[181] C. J. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," in *Proceedings of ICASSP*, pp. 384–386, Dallas, TX, 1987.

[182] P. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–47, 2002.

[183] P. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in *Proceedings of ICSLP'96*, pp. 1133–1136, Philadelphia, 1996.

[184] P. C. Woodland, "Hidden Markov models using vector linear predictors and discriminative output distributions," in *Proceedings of ICASSP*, San Francisco, 1992.

[185] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young, "The development of the 1996 HTK broadcast news transcription system," in *Proceedings of Human Language Technology Workshop*, 1997.

[186] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proceedings of ICASSP*, Toulouse, 2006.

[187] S. J. Young, "Generating multiple solutions from connected word DP recognition algorithms," in *Proceedings of IOA Autumn Conference*, vol. 6, pp. 351–354, 1984.

[188] S. J. Young and L. L. Chase, "Speech recognition evaluation: A review of the US CSR and LVCSR programmes," *Computer Speech and Language*, vol. 12, no. 4, pp. 263–279, 1998.

[189] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.4)*. University of Cambridge, `http://htk.eng.cam.ac.uk`, December 2006.

[190] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of Human Language Technology Workshop*, pp. 307–312, Plainsboro NJ, Morgan Kaufman Publishers Inc, 1994.

[191] S. J. Young, N. H. Russell, and J. H. S. Thornton, "Token passing: A simple conceptual model for connected speech recognition systems," Tech. Rep. CUED/F-INFENG/TR38, Cambridge University Engineering Department, 1989.

[192] S. J. Young, N. H. Russell, and J. H. S. Thornton, "The use of syntax and multiple alternatives in the VODIS voice operated database inquiry system," *Computer Speech and Language*, vol. 5, no. 1, pp. 65–80, 1991.

[193] K. Yu and M. J. F. Gales, "Discriminative cluster adaptive training," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, pp. 1694–1703, 2006.

[194] K. Yu and M. J. F. Gales, "Bayesian adaptive inference and adaptive training," *IEEE Transactions Speech and Audio Processing*, vol. 15, no. 6, pp. 1932–1943, August 2007.

[195] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised training with directed manual transcription for recognising Mandarin broadcast audio," in *Proceedings of InterSpeech*, Antwerp, 2007.

[196] H. Zen, K. Tokuda, and T. Kitamura, "A Viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features," in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[197] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *Proceedings of ICASSP*, Toulouse, 2006.

[198] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proceedings of InterSpeech*, Lisbon, Portugal, September 2005.

[199] Q. Zhu, Y. Chen, and N. Morgan, "On using MLP features in LVCSR," in *Proceedings of ICSLP*, Jeju, Korea, 2004.

[200] G. Zweig, *Speech Recognition with Dynamic Bayesian Networks*. PhD thesis, University of California, Berkeley, 1998.

[201] G. Zweig and M. Padmanabhan, "Boosting Gaussian Mixtures in an LVCSR System," in *Proceedings of ICASSP*, Istanbul, 2000.