The LIMSI 1999 Hub-4E Transcription System

Jean-Luc Gauvain, Lori Lamel, Gilles Adda,

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,lamel,gadda}@limsi.fr

ABSTRACT

In this paper we report on the LIMSI 1999 Hub-4E system for broadcast news transcription. The main difference from our previous broadcast news transcription system is that a new decoder was implemented to meet the 10xRT requirement. This single pass 4-gram dynamic network decoder is based on a time-synchronous Viterbi search with dynamic expansion of LM-state conditioned lexical trees, and with acoustic and language model lookaheads. The decoder can handle position-dependent, cross-word triphones and lexicons with contextual pronunciations. Faster than real-time decoding can be obtained using this decoder with a word error under 30%, running in less than 100 Mb of memory on widely available platforms such Pentium III or Alpha machines.

The same basic models (lexicon, acoustic models, language models) and partitioning procedure used in past systems have been used for this evaluation. The acoustic models were trained on about 150 hours of transcribed speech material. 65K word language models were obtained by interpolation of backoff n-gram language models trained on different text data sets. Prior to word decoding a maximum likelihood partitioning algorithm segments the data into homogenous regions and assigns gender, bandwidth and cluster labels to the speech segments. Word decoding is carried out in three steps, integrating cluster-based MLLR acoustic model adaptation. The final decoding step uses a 4-gram language model interpolated with a category trigram model. The overall word transcription error on the 1999 evaluation test data was 17.1% for the baseline 10X system.

1. INTRODUCTION

This paper describes the LIMSI 1999 broadcast news transcription system and reports on our development work prior to the fall 1999 Hub4 evaluation test. The baseline condition in this test imposed a computational time limit of 10 times real-time. In order to meet this requirement a new decoder was implemented which transcribes broadcast data in less than 10 times real-time with only a slight increase in word error rate when compared to our best system [10].

A major recent advance in speech recognition technology is the ability of todays systems to deal with non-homogeneous data as is exemplified by broadcast news: changing speakers, languages, backgrounds, topics. However transcribing such data requires significantly higher processing power than what is needed to transcribe read speech data in a controlled environment, such as for speaker adapted dictation. With the rapid expansion of different media sources for information dissemination, processing time is an important factor in making a speech transcription system viable for automatic indexation of radio and television broadcasts. A variety of near-term applications are possible such as audio data mining, selective dissemination of information, media monitoring services, disclosure of the information content and content-based indexation for digital libraries, etc. Current state-of-the-art laboratory systems can transcribe unrestricted broadcast news data with word error rates under 20%. When only concerned by the word error rate, it is common to design systems that run in 100 times real-time or more.

In designing a broadcast news transcription system with computational resources in the range of 10xRT, we compared performance using single pass or multiple pass decoding strategies. For each configuration the acoustic and language models were selected to optimize performance given the computational constraints. The influence of transcription accuracy on indexation performance was investigated using the TREC-8 SDR data [7, 10].

In the remainder of this paper we provide an overview of the LIMSI Nov99 Hub-4E system, with an emphasis on the new single pass decoder developed for this evaluation. Results are reported on a a representative portion of the Nov98 evaluation test set used for system development, as well as the 1999 evaluation test set. All the reported runs were done on a Compaq XP1000 500MHz machine with Digital Unix.

2. SYSTEM OVERVIEW

The LIMSI broadcast news automatic transcription system [3] consists of an audio partitioner [9], and a speech recognizer [4, 11].

The goal of audio partitioning is to divide the acoustic signal into homogeneous segments, labeling and structuring the acoustic content of the data. Partitioning consists of identifying and removing non-speech segments, and then clustering the speech segments and assigning bandwidth and gen-

der labels to each segment. The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels, which can be used to generate metadata annotations. While it is possible to transcribe the continuous stream of audio data without any prior segmentation, some of the advantages partitioning offers over such a straight-forward solution are given in [9].

The partitioning approach used in the LIMSI BN transcription system relies on an audio stream mixture model [9]. Each component audio source, representing a speaker in a particular background and channel condition, is in turn modeled by a GMM. The segment boundaries and labels are jointly identified by an iterative maximum likelihood segmentation/clustering procedure using GMMs and agglomerative clustering. The partitioning procedure (segmentation and labeling) is identical to the one used in the Nov'98 LIMSI HUB4E system [11], except for the number of iterations which is reduced to 8 for a slight speedup.

The partitioning procedure is as follows: First, the non-speech segments are detected (and rejected) using GMMs. Four GMMs each with 64 Gaussians serve to detect speech, pure-music and other (background). All test segments labeled as music or silence are removed prior to further processing. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. Given the sequence of cepstral vectors the algorithm tries to maximize an objective function defined as a penalized log-likelihood. Alternate Viterbi reestimation and agglomerative clustering gives a sequence of estimates with non-decreasing values of the objective function. The algorithm stops when no merge is possible. A constraint on the cluster size is used to ensure that each cluster corresponds to at least 10s of speech. This procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge, and the segment boundary penalty. When no more merges are possible, the segment boundaries are refined (within a 1s interval) using the last set of GMMs and an additional relative energy-based boundary penalty. This is done to locate the segment boundaries at silence portions, so as to avoid cutting words. Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to label telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with each word. [1] The speaker-independent large vocabulary, continuous speech recognizer makes use of n-gram statistics for language modeling and of continuous density HMMs with Gaussian mixtures for acoustic modeling. Word recognition is usually performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The hypotheses are used in cluster-based acoustic model adaptation using the MLLR technique [16] prior to word graph generation, and all subsequent decoding passes. The final hypothesis is generated using a 4-gram language model.

For all the experimental results given in this paper, the following training conditions were used. The acoustic models were trained on about 150 hours of American English broadcast news data. This data was used to train the Gaussian mixture models needed for segmentation and the acoustic models for use in word recognition. We used the August 1997 and February 1998 releases of the LDC transcriptions. Overlapping speech portions were detected in the transcriptions and removed from the training data. The phone models are position-dependent triphones, with about 11500 tied-states for the largest model set. Using word-position dependent triphone models, enables more accurate acoustic modeling at word boundaries as the contexts are limited to those triphones actually occurring in cross-word position. The state-tying is obtained via a divisive, decision tree based clustering algorithm with et of 184 questions concerning the distinctive features of the phone and the neighboring phones and the state positions. The number of triphone contexts and the amount of parameter sharing (state tying) influence the total model size (number of Gaussians) and consequently the decoding speed. Wideband and telephone band sets of gender-dependent acoustic models were built using MAP adaptation of SI seed models.

The acoustic analysis derives cepstral parameters from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10ms[6]. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficents and the log energy, along with the first and second order derivatives.

Fixed language models were obtained by interpolation of n-gram backoff language models trained on 3 different data sets: 203 M words of BN transcripts from LDC (years 92-

[1] Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to limit the memory required for the trigram and 4-gram decoding passes[6]. To do so a bimodal distribution is estimated by fitting a mixture of 2 Gaussians to the log-RMS power for all frames of the segment. This distribution is used to determine locations which are likely to correspond to pauses, thus being reasonable places to cut the segment. Cuts are made at the most probable pause 15s to 30s from the previous cut.

95) and from PSMedia (years 96, 97 and Jan/Feb'98); 343 M words of NAB newspaper texts and AP Wordstream texts (Jan'94 - Feb'98); 1.6 M words corresponding to the transcriptions of the acoustic training data (including all the dev and test sets predating Jan'98). The BN texts from PSmedia were processed using a modified version of the bn_raw2sgml.pl perl script from BBN made available by LDC. The broadcast news training texts were cleaned in order to be homogeneous with the previous texts, and filler words such as UH and UHM, were mapped to a unique form. All of the training texts (95 Hub3 and Hub4, and BN) were reprocessed to add a proportion of breath markers (4%), and of filler words (0.5%)[6]. The interpolation coefficients of these LMs were chosen so as to minimize the perplexity on the 2nd set of the Hub4 Nov98 evaluation data. The 4-gram LM contains 7M bigrams, 14M trigrams and 11M fourgrams.

The recognition word list contains 65122 words is identical to the one used in the Nov'98 LIMSI HUB-4E system [11], and has a lexical coverage of 99.7% and 99.5% on the Hub4-Nov98 set 2 and the eval99 test sets, respectively. The pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. Compound words are used for about 300 frequent word sequences and 1000 frequent acronyms [6].

3. SINGLE-PASS DECODER

A 4-gram single-pass dynamic network decoder has been developed. It is a time-synchronous Viterbi decoder with dynamic expansion of LM state conditioned lexical trees [1, 18, 17] with acoustic and language model lookaheads. The decoder can handle position-dependent, cross-word triphones and lexicons with contextual pronunciations. It makes use of various pruning techniques to reduce the search space and computation time, including three HMM-state pruning beams and fast Gaussian likelihood computations. It can also generate word graphs and rescore them with different acoustic and language models. Faster than real-time decoding can be obtained using this decoder with a word error under 30%, running in less than 100 Mb of memory on widely available platforms such Pentium III or Alpha machines.

The decoder by itself does not solve the problem of reducing the recognition time as proper models have to be used in order to optimize the recognizer accuracy at a given decoding speed. In general, better models have more parameters, and therefore require more computation. However, since the models are more accurate, it is often possible to use a tighter pruning level (thus reducing the computational load) without any loss in accuracy. Thus, limitations on the available computational resources can significantly affect the design of the acoustic and language models. For each operating point, the right balance between model complexity and pruning level must be found.

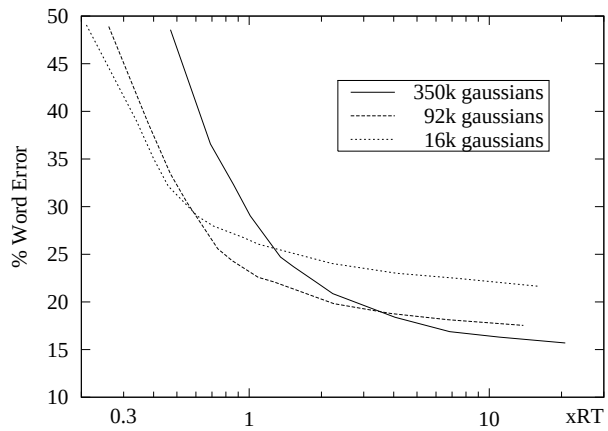

Figure 1: Word error rate vs. processing time for three acoustic model sets with 350k, 92k and 16k Gaussians on a subset of the Hub4-98 test data. (Single pass decoding with a trigram LM and no acoustic model adaptation.)

To illustrate this point, Figure 1 plots the word error rate as a function of processing time for 3 sets of acoustic models, which taken together minimize the word error rate over a wide range of processing times (from 0.3xRT to 20xRT) for the LIMSI broadcast news transcription system. It should be noted that transcribing such inhomogeneous data requires significantly higher processing power than for speaker adapted dictation systems, due to the lack of control of the recordings and linguistic content, which on average results in lower SNR ratios, a poorer fit of the acoustic and language models to the data, and as a consequence, the need for larger models. These results on a representative portion of the Hub4-98 eval test data are obtained using a 3-gram language model, and without acoustic model adaptation. The largest model set (350k Gaussians, 11k tied states, 30k phone contexts) provides the best performance/speed ratio for processing times over 5xRT. The 92k model set (92k Gaussians, 6k tied states, 5k phone contexts) performs better in the range of 0.6xRT to 3xRT, whereas a much smaller model set (16k Gaussians) is needed to go under real-time. Therefore depending upon the desired operating point different model set configurations will be most effective.

For a decoder based on lexical tree copies, the potential search space is proportional to the number of LM contexts, i.e., the number of n-1-grams in the backoff component of the n-gram LM. As observed for the acoustic models, there is a tradeoff between model complexity and search space, i.e. the best model without computational constraints may not be the best when such constraints are imposed. Figure 2 gives the word error rate as a function of the recognition time for four language models (1-gram to 4-gram LM) on the same representative subset of the Hub4-98 eval test data set. The same acoustic model set (6k states, 92k Gaussians) is used for all runs. It can be seen that the trigram LM is the best

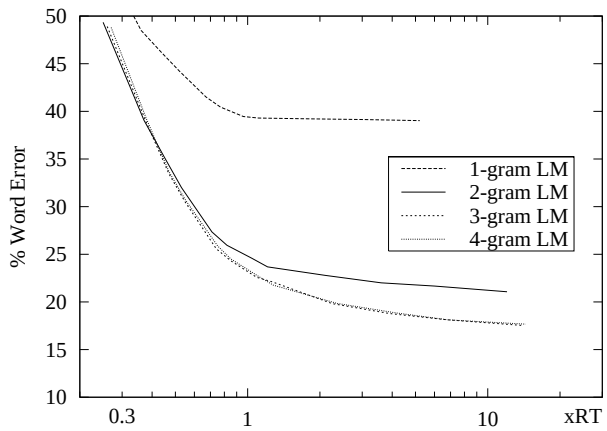

Figure 2: Word error rate vs. processing time for 4 language models (1-gram to 4-gram LM) on a subset of the Hub4-98 data. (Single pass decoding with the 92k acoustic model set and no adaptation.)

comprise for computation times in the range of interest (0.5 to 10xRT). In this range the 4-gram LM gives the same results, but requires about 50% more parameters than the 3-gram language model. The difference is even larger if the required memory space is compared. To observe a significant difference in favor of the 4-gram LM, the computation time needs to be over 20xRT with this single pass decoding. For computation times under 0.5xRT it does matter which LM order is used, as long as it is greater than 1.

4. MULTIPLE PASS DECODER

Many systems use a multiple pass decoding strategy to reduce the computational requirements. In multipass decoding, additional knowledge sources are progressively used in the decoding process, which allows the complexity of each individual decoding pass to be reduced and often results in a faster overall decoder. One of the main advantages of multiple pass decoding is the possibility to carry out acoustic model adaptation, such as unsupervised MLLR, between passes by making use of the current best hypotheses. Our targeted speed being lower than 10xRT, we need to pay attention to the computing resources required to perform the adaptation. In these experiments we use a single block diagonal regression matrix and run only one iteration of MLLR reestimation. Table 1 gives the computation time and word error rates for various decoding strategies. The pruning thresholds have been set so as to match the computing time of the most interesting setups. All passes perform a full decode, except the last decoding pass (labelled D) which is a word graph rescoring using a graph generated in the second 3-gram pass. The 3 acoustic model sets compared in Figure 1 are used, with the 16k Gaussian set used in the first pass, the 92k Gaussian set used in the second pass, and the 350k Gaussian set used in the last pass.

These results clearly demonstrate the advantage of using a multiple pass decoding approach. Comparing the setups A (1 pass, 6.8xRT, 16.8%) and D (2 passes, 6.9xRT, 15.4%),

	Pass	AM	*LM*	*RT*	*TotalxRT*	*Werr*
A	1	92k	3g	6.8	6.8	16.8%
B	1	350k	4g	10.5	10.5	16.1%
	1	92k	3g	0.8		24.7%
C	2	175k+mllr	4g	9.9	10.7	14.6%
	1	92k	3g	0.8		24.7%
D	2	175k+mllr	3g	6.1	6.9	15.4%
E	3	350k+mllr	4g	1.5	8.4	14.2%

Table 1: Comparison of decoding strategies on the NIST Hub4 eval98 set (partitioning and coding times are not included).

we see that the extra computing time needed for the first decode and the MLLR adaptation is largely compensated by the reduction in word error rate. Using adapted acoustic models allows us to use a tighter pruning threshold and have the same overall computing time but with a significantly lower word error rate. Also comparing setups C (2 passes, 10.7xRT, 14.6%) and E (3 passes, 8.4xRT, 14.2%) demonstrate the advantage of using an extra decoding pass to take advantage of the 4-gram LM and hypotheses for the MLLR adaptation.

As a result of these experiments, the configuration selected for the 1999 evaluation system has 3 decoding passes. The first pass generates initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via a one pass (1xRT) cross-word trigram decoding with gender-specific sets of position-dependent triphones (5400 contexts and 6275 tied states) and a trigram language model (17M trigrams and 8M bigrams). Band-limited acoustic models are used for the telephone speech segments. Prior to the second pass, which generates a word graph, unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [16] A word graph is generated for each segment in a one pass (about 6xRT) trigram decoding using position-dependent triphones covering 28k contexts with 11700 tied states (16 Gaussians per state) and the trigram used in the first pass. The final hypothesis is generated after a second MLLR adaptation using the word graphs, a 4-gram model and a 32-Gaussian version of the acoustic models used in pass 2. These third pass model sets are quite comparable in size to that used in our 1998 system (covering 28k phone contexts with 11500 tied states). Band-limited versions of the acoustic models are used for the telephone speech segments.

In Table 2 the word error rates and the total computation time (including partitioning) are given for both the development test set (Hub4 eval98) and the Hub4 eval99 test set. For reference, the official result on the eval98 test set using our Nov98 system was 13.6%, with a decoding time around 200xRT [11]. Using only the first decoding pass, unrestricted BN data can be decoded in less than 1.4xRT (including partitioning) with a word error rate around 30%.

A 10xRT contrast system was also developed, which used

	Dev data (eval98)		*Test data (eval99)*	
Step	*CPU time*	*Werr*	*CPU time*	*Werr*
Coding and Partitioning:	0.5xRT		0.5xRT	
Word decoding:				
pass#1 (generate 3-gram hyp):	0.8xRT	24.7%	0.9xRT	29.3%
pass#2 (MLLR, 3-gram):	6.1xRT	15.4%	6.5xRT	18.5%
pass#3 (MLLR, 4-gram):	1.5xRT	14.2%	1.5xRT	17.1%
Overall:	8.9xRT	14.2%	9.4xRT	17.1%

Table 2: 10xRT results in word error rate for the NIST BN 1998 and 1999 test sets.

the same decoding strategy but made use of additional acoustic and language model training data from the TDT-2 corpus. The recognition word list for this contrast system contained 65343 words (with 77033 pronunciations), of which ∼3800 were not in the baseline system word list. The third decoding pass made use of acoustic models from our 1998 system which were adapted (via MAP adaptation) with about 500 hours of TDT-2 acoustic data from February-May 1998. Since no detailed transcriptions were available for this data, only segments for which the word hypotheses matched the closed caption were used for training. Additional training texts from the period of March through May 1998 (from PS-Media, newswires) and the TDT-2 closed captions and commercial transcripts (predating June98) were also used to estimate the language models. The additional data gave a slight performance improvement (3% relative) on the Nov'98 evaluation data, but only a 0.1% absolute reduction to 17.0% on the 1999 test set.

An unconstrained computation system was also evaluated, in which the 10x baseline result served as the initial hypotheses for further decoding. The acoustic and language models were the same as those used for the 10xRT baseline. Two additional decoding passes were carried out: word graph generation with MLLR adapted acoustic models and a 4-gram language model (32xRT) and word graph rescoring with MLLR adapted acoustic models a 4-gram language model (5xRT). The results for this system are given in Table 3. Unfortunately the version of the decoder script used for the evaluation run had a bug which both made it run slower and had a higher word error rate. The reduction in word error is less than 10% compared to the baseline 10x systems despite the factor of 5 in computation time. Some recent experiments with Rover [19] support previous observations that combining even a small number of fast decoders may be more efficient in reducing the recognition word error rate than running a slower system.

5. SUMMARY & DISCUSSION

In this paper we have presented our 1999 broadcast news transcription system, and highlighted our development work which was mainly oriented towards developing a new decoder and optimizing the acoustic and language models to remain within the 10x computational restriction. With this

	Submitted (bug)	*Corrected script*
xRT	54xRT	47xRT
bn99en-1	17.4%	17.0%
bn99en-2	14.8%	14.5%
average	15.9%	15.6%

Table 3: 50xR contrast system.

competitive new decoder unrestricted broadcast news data can be transcribed in under 1.4xRT with a word error under 30%. Different decoding strategies were investigated so as to optimize performance at for different real-time factors. A three pass strategy was found to provide the best performance at 10xRT, whereas fewer passes are better for faster decoding speeds.

Our development work showed us how processing time constraints significantly affect model design. For each operating point, the right balance between model complexity and search pruning level must be found. For moderate decoding times (in the range 0.6xRT to 3xRT) a model set containing 92k Gaussians, 6k tied states, 5k phone contexts was found to perform substantially better than smaller or larger models. If decoding time is not an important factor, an even larger model set (350k Gaussians, 11k tied states, 30k phone contexts) provided the best performance/speed ratio for processing times over 5xRT.

We have developed broadcast news transcription systems with comparable performance levels for three other languages[2] (French, German and Mandarin) and are currently also targeting Portuguese and Arabic.

REFERENCES

[1] X. Aubert, "One Pass Cross Word Decoding for Large Vocabularies Based on a Lexical Tree Search Organization," *Proc. ESCA Eurospeech'99*, **4**, pp. 1559-1562, Budapest, Hungary, September 1999.

[2] S.S. Chen, P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *DARPA Broadcast News Transcription & Understanding Workshop*, pp. 127-132, Feb. 1998.

[2] The French and German systems have been partially supported by the EC Olive project and the French Ministry of Defense (DGA). The data were provided by INA and ARTE, partners in the Olive project, as well as the DGA.

[3] J.L. Gauvain, L. Lamel, and G. Adda, 'Transcribing broadcast news for audio and video indexing," Communications of the ACM, 43(2), Feb 2000.

[4] J.L. Gauvain, L. Lamel, G. Adda, "Recent Advances in Transcribing Television and Radio Broadcasts," *Proc. Eurospeech'99*, **2**, pp. 655-658, Budapest, Sept. 1999.

[5] J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, **2**(2), pp. 291-298, April 1994.

[6] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *ARPA Speech Recognition Workshop*, Chantilly, pp. 56-63, Feb. 1997.

[7] J.L. Gauvain, Y. de Kercadio, L.F. Lamel, G. Adda "The LIMSI SDR system for TREC-8," *Proc. of the 8th Text Retrieval Conference TREC-8*, pp. 405-412, Gaithersburg, MD, November 1999.

[8] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI 1997 Hub-4E Transcription System", *Proc. DARPA Broadcast News Transcription & Understanding Workshop*, pp. 75-79, Feb. 1998.

[9] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, **5**, pp. 1335-1338, Sydney, Dec. 1998.

[10] J.L. Gauvain, L. Lamel, G. Adda, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP'00*, **5**, pp. 1335-1338, Beijing, Oct. 2000.

[11] J.L. Gauvain, L. Lamel, G. Adda, M. Jardino, "The LIMSI 1998 Hub-4E Transcription System," *Proc. DARPA Broadcast News Workshop*, pp. 99-104, Feb. 1999.

[12] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," *ICASSP-91*, pp. 873-876, May 1991.

[13] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, S.J. Young, "Segment Generation and Clustering in the HTK Broadcast News Transcription System," *DARPA Broadcast News Transcription & Understanding Workshop*, pp. 133-137, Feb. 1998.

[14] M. Jardino "Multilingual stochastic n-gram class language models," *ICASSP-96*, Atlanta, May 1996.

[15] A. Kannan, M. Ostendorf, J.R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," *IEEE Trans. Speech & Audio*, **2**(3), July 1994.

[16] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

[17] H. Ney, R. Haeb-Umbach, B.H. Tran and M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE ICASSP-92*, **I**, pp. 9-12, San Francisco, CA, March 1992.

[18] J.J. Odell, V. Valtchev, P.C. Woodland and S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proc. ARPA Human Language Technology Workshop*, pp. 405-410, Princeton, NJ, March 1994.

[19] H. Schwenk, J.L. Gauvain, "Improved Rover using Language Model Information," *Proc. ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, pp. 47-52, Paris, Sep 2000.

[20] M. Siegler, U. Jain, B. Raj, R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *DARPA Speech Recognition Workshop*, Chantilly, pp. 97-99, Feb. 1997.

[21] P.C. Woodland, T. Neieler, E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR", presented at the 1998 Hub5E Workshop, Sept. 1998.