DESIGN OF FAST LVCSR SYSTEMS

G. Evermann & P.C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {ge204,pcw}@eng.cam.ac.uk

ABSTRACT

This paper describes the development of fast (less than 10 times real-time) large vocabulary continuous speech recognition (LVCSR) systems based on technology developed for unlimited runtime systems assembled for participation in recent DARPA/NIST LVCSR evaluations. A general system structure for 10 times real-time systems is proposed and two specific systems that have been built for Broadcast News (BN) and Conversational Telephone Speech (CTS) recognition are described. The systems were evaluated in the DARPA/NIST April 2003 Rich Transcription evaluation. Results are reported and contrasted with unlimited runtime systems and previous fast systems.

1. INTRODUCTION

For more than a decade a major focus in LVCSR research work have been the yearly U.S. Government sponsored evaluations conducted by NIST. While these evaluations helped the research community to accurately measure the progress in the state-of-the-art in LVCSR and led to impressive improvements in accuracy [17], they also encouraged research sites to pursue "accuracy at any price". This lead to typical systems running in about 300 times slower than real-time (with some taking up to 2000xRT). As LVCSR technology matured there is now again an increased interest in building faster systems while retaining the gains achieved. This trend is also reflected in the recently initiated DARPA EARS programme which aims at fast transcription of both Broadcast News and Conversational Telephone Speech data.

Building very fast systems (faster than real time) on difficult tasks like CTS typically involves sacrificing many of the advanced techniques that have been developed in recent years and also requires a significant amount of specific low-level software optimisation which is not necessarily useful for general research use. Two impressive examples of systems that can run in real-time on the CTS task are the 2002 AT&T system [12] and the 2003 IBM system [13]. Due to the runtime restrictions both of these lack a number of important features typically found in larger LVCSR evaluation systems and thus have significantly higher word error rates than the systems discussed in this paper.

However, systems in the range of about 10xRT can be built based on existing research LVCSR technology if the system is designed carefully while employing most of the standard state-of-the-art modelling techniques. This paper discusses the issues involved in developing systems that run in less than 10xRT. Two such systems were developed at Cambridge and entered in the April 2003 Rich Transcription evaluation for Broadcast News (BN) and Conversational Telephone Speech (CTS) respectively [1].

The structure of the paper is as follows. First a short overview of the recognition tasks considered and the available training data is given. In section 3 the sets of models (both acoustic and language models) trained for the two tasks are described with particular emphasis on the common techniques used for both tasks.

In section 4 a general system structure for fast LVCSR systems is proposed and in the following section a quick overview of previous fast systems developed at CUED is given. The next two sections discuss specifics of the actual 2003 BN and CTS systems. Section 8 describes experiments related to ensuring that the systems ran in less than 10xRT. An overview of the performance of the 2003 systems is given in section 9 contrasting it with previous years' systems, both fast and otherwise. The paper concludes with a discussion of directions for future work.

2. TASKS

The two most commonly used tasks for LVCSR research in English are Broadcast News (BN, formerly known as Hub-4) and Conversational Telephone Speech (CTS, often referred to as Switchboard or Hub-5). For BN the test data is taken from radio and TV news shows (for example CNN Headline News, ABC World News Tonight). For CTS the test data consists of excerpts of telephone conversations conducted by volunteers on an assigned topic. Both tasks were used in the DARPA sponsored April 2003 Rich Transcription evaluation conducted by NIST.

The acoustic training data available for BN consists of

To appear in *Proc. ASRU 2003*

143 hours released by the LDC in 1997 and 1998. BN LMs were trained on the acoustic transcripts (2 million words), a number of other broadcast news transcript sources (343M words) plus a variety of newspaper texts (674M words). More details on the task are given in [4]. The 2003 BN eval test set consists of 6 half-hour excerpts from Radio and TV broadcasts taken from February 2001. A development set (dev03) with similar properties was selected and transcribed (see [8] for details).

For CTS more acoustic training data is available: 296 hours released by LDC (Switchboard I, Call Home English and Switchboard Cellular) plus an additional 67 hours of Switchboard (Cellular and Switchboard II phase 2). For the LDC data detailed, careful transcriptions are available. For the additional 67 hours BBN made "quick transcriptions" available that were produced by a commercial transcription service. The CTS LM is trained on the acoustic transcripts plus parts of the BN LM data. For the 2003 evaluation data was taken from the new LDC Fisher collection[1] and Switchboard II phase 5. The set contains 72 excerpts of 5 minutes each for a total of about 6 hours. The conversations were chosen to balance gender and to contain a mix of landline and cellular calls (3:1 for Fisher and 1:1 for Switchboard). Two sets were used for the development (dev01, eval02). They are similar to eval03 in composition.

3. MODELS

The acoustic and language models for the two tasks were built using a similar set of procedures and techniques. The common techniques include:

- The audio file is parameterised using 13 PLP features augmented with their first, second and third derivatives normalised by Cepstral Mean Normalisation (CMN).
- The resulting 52 dimensional features are projected down to 39 using a global Heteroscedastic Linear Discriminant Analysis (HLDA) transform. See [9] for details of the transform estimation.
- Acoustic models are trained using the Minimum Phone Error criterion [10].
- A separate model-set was trained using Speaker Adaptive Training (SAT) employing constrained MLLR.
- The base dictionary contained about 1.1 pronunciations per word on average. The pronunciation dictionary was originally based on the 1993 LIMSI WSJ lexicon and phone set but many words have been added or modified. A special single pronunciations (SPron) versions was created and used to trained a separate HMM set. [5].
- During recognition these models are adapted with multiple full-matrix linear mean transforms, diagonal variance transforms and a global full-variance transform all estimated using lattice MLLR [14]
- The language model used an interpolation of a word-based fourgram and a class-based trigram with automatically derived classes.

3.1. BN specific modelling

For BN it has been shown in the past that it is advantageous to employ bandwidth-specific, gender-dependent acoustic models [15]. Therefore narrow-band models were trained on bandpass-filtered version of the training data. Gender-specific models were derived from the gender-independent models using MPE-MAP [11], which allows using MAP adaptation while retaining the advantage of discriminative training. More details on the effectiveness of these techniques on Broadcast News can be found in [8].

3.2. CTS specific modelling

The CTS system employed Vocal Tract Length Normalisation (VTLN) which was applied both in training and test by warping the filterbank. After applying VTLN and CMN the variance of the features was normalised on a conversation side basis (CVN). An in-depth discussion of the issues involved in recognising conversational telephone speech and past systems developed at CUED is given in [6].

4. GENERAL SYSTEM STRUCTURE

All state-of-the-art unrestricted compute LVCSR evaluation systems developed in recent years run in multiple passes and use system combination to derive the final output, based on techniques like ROVER [3]. To make the use of multiple model sets feasible, these systems typically employ lattices to restrict the search space for the later stages of the system. The system structure proposed here consists of two main stages: the initial lattice generation stage and the rescoring stage in which these lattices are rescored which multiple model sets.

The purpose of the lattice generation stage is two-fold. The resulting lattices are used to restrict the search space for the rescoring stage and the 1-best word sequence produced in the lattice generation stage is used as the adaptation supervision for each of the rescoring model sets.

Figure 1 shows the proposed system structure. The first step is to segment the audio stream into speech segments of manageable size (e.g. up to 30 seconds in length) discarding all non-speech portions of the audio (silence, music, etc.).

[1] http://www.ldc.upenn.edu/Fisher/

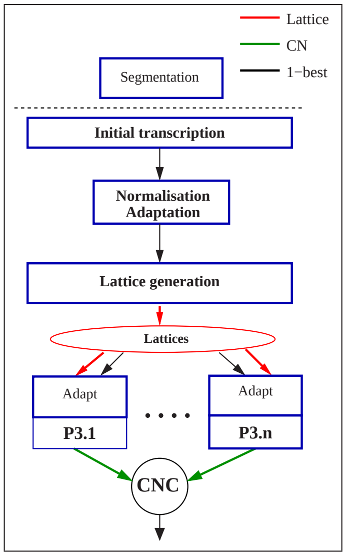

Fig. 1. General System Structure

The next step is the first decoding pass (P1) which generates an initial transcription that will then be used for normalisation/adaptation purposes. Typical examples are Vocal Tract Length Normalisation or global MLLR adaptation. The word boundary times of the initial transcription can also be used to improve the initial segmentation. The accuracy of the transcription is not that critical as transform-based adaptation techniques with a small number of free parameters are fairly robust with regard to the supervision quality and since no hard decisions regarding the search space are taken. Thus the initial transcription can be performed with simplified models and with very tight pruning settings.

The next decoding pass (P2) is the lattice generation using adapted detailed acoustic models and the best LM available. The resulting lattices together with their 1-best transcriptions form the input to the multiple rescoring passes.

In the rescoring stage the processing is split into multiple *branches*. In each the lattices are rescored with a different set of acoustic models. These rescoring results are then combined to yield the final system output. The more model sets are available for rescoring passes (P3.1 – P3.n) the better, but the overall runtime constraint might limit the number of passes that can be run. When only a small number of branches can be used it is important to carefully choose the model-sets to maximise the gain from combination.

For each of the available model-sets the word hypotheses generated in the lattice generation stage are used to adapt the model-set to the current speaker and environment conditions. With these adapted models the rescoring lattices are re-decoded and new output lattices are produced which are then converted into confusion networks.

The resulting confusion networks (one per model-set for each segment) are then combined using Confusion Network Combination (CNC, see [2]) to yield the final system output with associated confidence scores.

A typical compute time budget for a system with two rescoring model sets (branches) is shown in Table 1. These times were used as guidelines during system development.

1xRT	Initial Transcription
0.5xRT	Normalisation/Adaptation
4xRT	Lattice Generation
2xRT	adapt/rescore (per branch)

Table 1. Typical Compute Budget for 2 Branch System

5. PREVIOUS WORK ON FAST SYSTEMS

Over the last decade many different LVCSR systems were developed at CUED.[2] These systems generally increased in complexity over the years. The Hub-5 systems developed in the past few years ran between 200 and 300xRT and employed 6 to 8 rescoring branches for system combination.

In 1998 a Broadcast News system was developed jointly by Entropic and CUED that ran in less than 10xRT. This system ran in two passes (comparable to the “Initial Transcription” and “Lattice Generation” passes in Figure 1). It performed automatic segmentation and employed speaker clustering and adaptation. For a detailed description see [15]. The word error rate of this system was 16.6% relative higher than for the full (300xRT) system (16.1% vs. 13.8% on bneval98).

For the 2002 Hub-5 evaluation a fast version of the full (320xRT) system was developed. This system made use of the triphone models built for the large system. It employed a simplified system structure and much tighter pruning settings to speed up the search in all passes. The system used three passes as shown in Figure 1, but only a single branch in the rescoring stage, i.e. no system combination. Details can be found in [16]. The error rate of the fast system was 13.8% relative higher than for the full system (27.2% vs. 23.9% on eval02).

6. 2003 CTS SYSTEM

Based on the available compute platform[3] and the experience with the 2002 CTS system it was decided to aim for two branches in the P3 rescoring stage of the 2003 system.

Four different kinds of triphone model sets were trained for the unlimited compute 2003 CTS system. They all em-

[2] for an overview see `http://htk.eng.cam.ac.uk/docs/cuhtk.shtml`

[3] IBM x335 servers with 2.8GHz Intel Xeon CPUs, 512KB cache, 400MHz bus

ployed MPE training, but differed in the set of other techniques used:

A: SAT HLDA **B:** HLDA
C: SPron HLDA **D:** non-HLDA

The performance of these models was investigated in the framework of the unlimited compute CU-HTK system. Large lattices were generated with model set B and rescored independently with all four model sets (adapted using lattice MLLR and full-variance transforms). All decoding passes were run at very conservative beamwidths. Confusion network decoding was applied to the resulting four sets of lattices. The WER of each of the four model sets is shown in the first row of Table 2. The result of pairwise system combination using CNC is given in the rest of the table.

System	A	B	C	D
	23.0	23.6	23.4	24.8
+A		23.1	**22.6**	22.7
+B			22.9	23.3
+C				22.8

Table 2. Individual Systems and Pairwise Combination, %WER on eval02 after lattice MLLR/FV and CN

The combination of the SAT and the SPron models gave the best performance and thus these two models were chosen for the P3 rescoring stage in the 10xRT system. For the P2 lattice generation stage model set B was chosen to avoid biasing the lattices towards either of the P3 models. The resulting system structure is shown in Figure 2

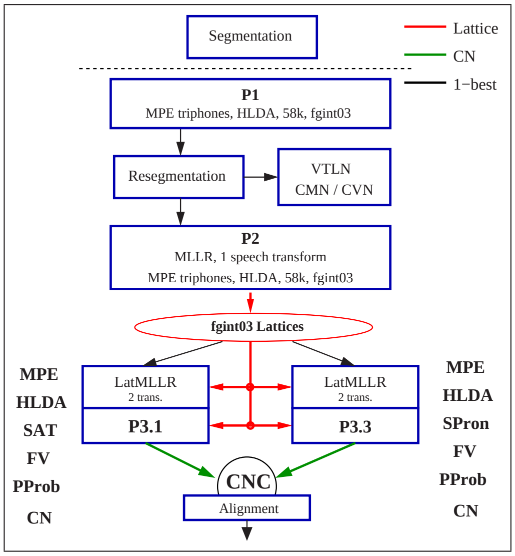

Fig. 2. CTS System Structure

7. 2003 BN SYSTEM

The BN system was built in a very similar fashion to the CTS system. Only a SAT and a SPron model set were built for the P3 stage. The P2 stage again used an HLDA MPron MPE model. Due to time constraints the SAT model was only trained on the wideband data whereas four versions of the P2 and the SPron model were built (male/female and wide-/narrow-band). To compensate for the lack of a narrowband SAT model the output of the P2 stage was also used in the system combination, leading to 2-way combination for narrowband data and 3-way combination for wideband data. The system employed automatic gender/bandwidth classification and speaker clustering based on the approach presented in [7]. The system structure is shown in Figure 3.

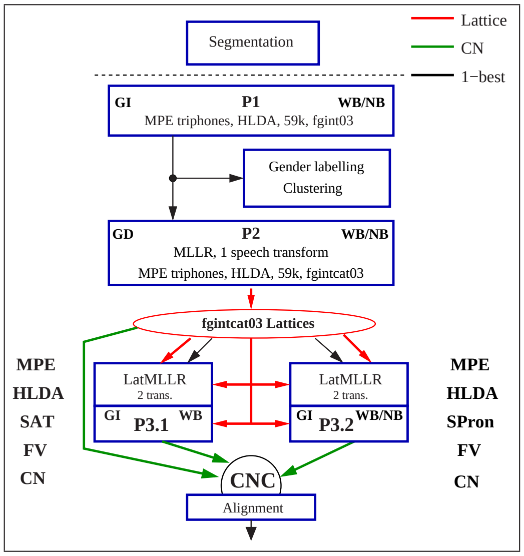

Fig. 3. BN System Structure

8. CONTROLLING RUNTIME BEHAVIOUR

To achieve a run time of less than 10 times real time it was necessary to run all parts of the system at operation points quite different from the setup normally used in full (200+ xRT) systems. The compute budget show in Table 1 was used as a starting point for tuning the system. In an effort to stay inside this compute budget initial parameter settings were chosen based on prior experience with the decoding setups used.

An initial experiment on the CTS eval02 test data (with manual segmentation) was run to choose an operating point for the initial transcription pass (P1). The first two passes of the system were run a few times while varying the P1 setup. The only significant influence that the P1 transcription has on the system performance is by serving as the adaptation supervision for P2. Table 3 shows that adaptation is relatively robust to changes in P1 WER. The middle operating point was chosen for all further work.

In the system structure used here a vital factor is the P2 lattice generation pass as it consumes the most time and directly affects the speed of the P3 rescoring passes.

P1 speed	WER		
xRT	P1	P2 trigram	P2 fourgram
0.48	37.4	26.3	25.5
0.83	35.2	26.3	25.4
1.50	34.4	26.1	25.2

Table 3. P1 speed-accuracy trade-off (CTS eval02)

Experiments confirmed that the time needed for rescoring the lattices can be accurately predicted from the lattice sizes. The time is proportional to the number of nodes in the tree rescoring network. The rescoring network size in turn grows roughly logarithmically with the lattice density.

The P2 lattice generation operating point was chosen so that the decoder ran in about 3xRT. The resulting lattices were then pruned so that the lattice rescoring could run in about 1xRT. On eval02 the fourgram lattices had a lattice (oracle) word error rate of 7.4% at a lattice density of 282 (number of arcs per reference word). Table 4 shows the run-times for all passes on the CTS eval03 set.

pass	run time
Coding + Segmentation	0.068 xRT
Pass1	0.890 xRT
VTLN	0.387 xRT
Pass2	3.178 xRT
Pass3.1 Adapt	1.422 xRT
Pass3.1 Rescore + CN	1.074 xRT
Pass3.2 Adapt	1.200 xRT
Pass3.2 Rescore + CN	0.919 xRT
Sys-Combine + align	0.068 xRT
total	9.207 xRT

Table 4. Run-times on CTS eval03

9. RESULTS

The 10xRT CTS system described above was evaluated in the 2003 DARPA/NIST Rich Transcription evaluation. The breakdown of the performance by stage is given in Table 5. For comparison the performance of the full (190xRT) CU-HTK system, which used 6-way system combination (3 triphone + 3 quinphone systems), is indicated in the first line.

It can be seen that the SAT and SPron models give very similar performance (0.2% absolute difference), but nevertheless their hypotheses are sufficiently different that their combination yields a further gain of 0.4% absolute over the best single model.

Table 6 shows the performance difference between the full systems and the fast systems. In the 2003 systems the performance difference is small at 1.6% absolute, despite the fact that the fast system lacks quinphones, uses only two triphone models and operates with a much more tightly pruned search. Comparing the 2003 fast system with the 2002 one, it can be seen that the use of more complex adaptation (lattice MLLR+FV) and the use of system combination brought the performance of the fast systems much closer to the performance of the respective full (6-way combination) systems[4]. The relative gap in 2002 was 14% and thus comparable to the previous BN 10xRT work discussed in section 5. In 2003 the gap has been reduced to 7%.

	Swbd II-5	Fisher	Total
190xRT	24.1	17.1	20.7
P1	39.0	29.7	34.5
P2	29.4	20.9	25.3
P3.1-cn	26.0	18.8	22.5
P3.3-cn	26.3	18.9	22.7
final	25.5	18.4	22.1

Table 5. CTS 2003 10xRT system, %WER on eval03

system	Swbd I	Swbd II-3	Cell	Total	gap
full ’02	19.8	24.3	27.0	23.9	
10x ’02	22.3	27.7	31.0	27.2	+14%
full ’03	18.6	22.3	23.7	21.7	
10x ’03	19.9	23.5	25.8	23.3	+7%

Table 6. CTS fast-gap, %WER on eval02 for full/fast systems (2002 systems used manual segmentation)

The overall progress in the state-of-the-art in LVCSR on the CTS task is documented in Table 7 which show the performance of the CU-HTK evaluation systems developed since 2000 tested on the dev01 test set. Steady progress has been made over the last years and fast systems now produce very competitive performance and can thus be used for system development or even application deployment.

year	xRT	Swbd I	Swbd II-2	Cell	Total
2000	255	19.3	32.5	33.2	28.3
2001	190	18.3	31.9	32.1	27.3
2002	320	16.4	29.2	27.4	24.2
2002	10	18.3	31.9	31.0	27.0
2003	10	15.8	26.9	25.9	22.8

Table 7. Progress on CTS, %WER on dev01 over the years

On both the eval02 and the dev01 test sets the 2003 fast system gives about 15% relative lower word error rates than the 2002 fast system.

[4] However, it has to be noted that the two fast systems were run on different computers, i.e. the 2002 system would now be faster than 10xRT

The performance of the BN system in the 2003 Rich Transcription evaluation for both the development set (dev03) and the official eval set (eval03) are shown in Table 8.

The behaviour on the two test sets is surprisingly different. On dev03 the rescoring with more complex adaptation and better models in the P3 stage yields significant improvements over the result of the P2 stage (0.8% abs.), but on eval03 this improvement is much smaller (0.2% abs.). However, the 3-way system combination is more effective on the eval03 (0.7% abs. vs. 0.4%). This indicates that the use of system combination increases the robustness of the overall system against variability in the test data.

	dev03	eval03
P1	15.9	14.6
P2.fgintcat	13.1	11.9
P2.fgintcat-cn	12.8	11.6
P3.1-cn† (SAT)	12.0	11.4
P3.3-cn (SPron)	12.1	11.4
final	11.6	10.7

Table 8. BN 2003 10xRT results,%WER on dev03/eval03, † wideband only, narrowband from P3.3

10. CONCLUSIONS & FUTURE WORK

In this paper a general structure for 10 times real time LVCSR systems was proposed. The specifics of conversational telephone speech and broadcast news systems were discussed and experimental results on the 2003 Rich Transcription evaluation task were presented. The CU-HTK system was the only system entered in the CTS 10xRT category. For CTS an improvement in WER of about 15% relative was achieved compared to last year's 10xRT system and the gap between the fast system and the full (200+ xRT) system was decreased significantly from 14% to 7%.

The BN system proved to be very robust against variability in the test data and achieved the lowest word error rate in the 10xRT BN category in the 2003 RT evaluation.

The 10xRT systems provide a good basis both for the investigation of new modelling techniques (both for acoustics and LMs) as their effectiveness can be easily and quickly tested in the context of a realistic full system. They can also be used in research scenarios that require large scale transcription of audio data, such as the investigation of lightly supervised training approaches. The proposed general system structure provides a framework for the development of both research systems and actual deployed products.

ACKNOWLEDGMENTS

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. The authors would like to thank all members of the HTK STT group, in particular T. Hain, X. Liu, L. Wang, D. Povey, D.Y. Kim, M.J.F. Gales, H.Y. Chan, D. Mrva, S.E. Tranter, K. Yu.

11. REFERENCES

[1] G. Evermann, D.Y. Kim, L. Wang, and P.C. Woodland. CU-HTK Fast System Description. In *Proc. Rich Transcription Workshop*, 2003.

[2] G. Evermann and P.C. Woodland. Posterior Probability Decoding, Confidence Estimation and System Combination. In *Proc. Speech Transcription Workshop*, 2000.

[3] J.G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In *Proc. IEEE ASRU Workshop*, 1997.

[4] D. Graff. An Overview of Broadcast News Corpora. *Speech Communication*, 37:15–26, 2002.

[5] T. Hain. Implicit Pronunciation Modelling in ASR. In *Proc. ISCA ITRW PMLA*, 2002.

[6] T. Hain, P.C. Woodland, G. Evermann, M.J.F. Gales, X. Liu, G.L. Moore, D. Povey, L. Wang. Automatic Transcription of Conversational Telephone Speech. Technical Report CUED/F-INFENG/TR 465, Cambridge University Engineering Department, 2003.

[7] S.E. Johnson and P.C. Woodland. Speaker Clustering Using Direct Maximisation of the MLLR-adapted Likelihood. In *Proc. ICSLP*, 1998.

[8] D.Y. Kim, G. Evermann, T. Hain, D. Mrva, S.E. Tranter, L. Wang, P.C. Woodland. Recent Advances in Broadcast News Transcription *to appear Proc. ASRU*, 2003.

[9] X. Liu and M.J.F. Gales. Automatic Complexity Control for HLDA Systems. In *Proc. ICASSP*, 2003.

[10] D. Povey and P.C. Woodland. Minimum Phone Error and I-Smoothing for Improved Discriminative Training. In *Proc. ICASSP*, 2002.

[11] D. Povey, P.C. Woodland, and M.J.F. Gales. Discriminative MAP for Acoustic Model Adaptation. In *Proc. ICASSP*, 2003.

[12] M. Riley, E. Bocchieri, A. Ljolje, and M. Saraclar. The AT&T 1x Real-time Switchboard Speech-to-text System. In *Proc. Rich Transcription Workshop*, 2002.

[13] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari. An Architecture for Rapid Decoding of Large Vocabulary Conversational Speech. In *Proc. Eurospeech*, 2003.

[14] L.F. Uebel and P.C. Woodland. Speaker Adaptation Using Lattice-based MLLR. In *Proc. ISCA ITRW on Adaptation Methods in Speech Recognition*, 2001.

[15] P.C. Woodland. The Development of the HTK Broadcast News Transcription System: An Overview. *Speech Communication*, 37, 2002.

[16] P.C. Woodland, G. Evermann, M.J.F. Gales, T. Hain, X. Liu, G.L. Moore, D. Povey, and L. Wang. CU-HTK April 2002 Switchboard System. In *Proc. Rich Transcription Workshop*, 2002.

[17] S.J. Young and L.L. Chase. Speech Recognition Evaluation: A Review of the US CSR and LVCSR Programmes. *Computer Speech & Language*, 12(4):263–279, 1998.