DEVELOPMENT OF THE 2003 CU-HTK CONVERSATIONAL TELEPHONE SPEECH TRANSCRIPTION SYSTEM

G. Evermann, H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, P.C. Woodland

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

ABSTRACT

This paper describes the development of the 2003 CU-HTK large vocabulary speech recognition system for Conversational Telephone Speech (CTS). The system was designed based on a multi-pass, multi-branch structure where the output of all branches is combined using system combination. A number of advanced modelling techniques such as Speaker Adaptive Training, Heteroscedastic Linear Discriminant Analysis, Minimum Phone Error estimation and specially constructed Single Pronunciation dictionaries were employed. The effectiveness of each of these techniques and their potential contribution to the result of system combination was evaluated in the framework of a state-of-the-art LVCSR system with sophisticated adaptation. The final 2003 CU-HTK CTS system constructed from some of these models is described and its performance on the DARPA/NIST 2003 Rich Transcription (RT-03) evaluation test set is discussed.

1. INTRODUCTION

Despite many years of research the transcription of Conversational Telephone Speech (CTS) remains one of the most challenging automatic speech recognition tasks. As a result of the difficulty of the problem and the intense research effort, highly complex recognition systems have been constructed and are evaluated in the annual U.S. government sponsored evaluations

In this paper a number of acoustic modelling techniques such as Speaker Adaptive Training (SAT), Heteroscedastic Linear Discriminant Analysis (HLDA), Minimum Phone Error (MPE) estimation and specially constructed Single Pronunciation dictionaries (SPron) are investigated in the framework of a state-of-the-art multi-pass CTS transcription system with sophisticated adaptation, large-scale language models and confusion network based system combination.

The rest of the paper is organised as follows. Section 2 gives an overview of the CTS task, properties of the training data and the basic features of the CU-HTK system. In Section 3 the theory of system combination is described. Section 4 describes the overall structure of the multi-branch CU-HTK CTS system and discusses the role of system combination. The next two sections present results regarding the performance of the advanced modelling technique used (SAT, HLDA, SPron) and their effect on the output of the system combination, respectively. Finally the performance of the complete 2003 CU-HTK system in the RT-03 evaluation is analysed.

2. BASIC SYSTEM FEATURES

The CTS data consists of phone conversations between by volunteers on an assigned topic in (American) English. The data available for training the acoustic models consists of 296 hours of speech released by the LDC (Switchboard I, Call Home English and Switchboard Cellular) plus 67 hours of Switchboard (Cellular and Switchboard II phase 2). For the LDC data detailed, careful transcriptions were provided by MSState University. For the additional 67 hours BBN made "quick transcriptions" available that were produced by a commercial transcription service.

A word-based 4-gram language model was trained on the acoustic transcriptions, additional Broadcast News data (427M words of text) plus 62M words of "conversational texts" collected from the World Wide Web [1]. The word-based 4-gram was smoothed with a class-based trigram trained only on the CTS transcriptions. The recognition dictionary contained 58,231 words with an average of 1.1 pronunciations per word.

The evaluation test set used in the 2003 DARPA/NIST Rich Transcription evaluation (*eval03*) contains data from the LDC Fisher collection[1] and from Switchboard II phase 5. The set comprises 72 phone calls of 5 minutes each for a total of about 6 hours chosen to balance gender. It contains a mix of landline and cellular calls. An additional DARPA/NIST internal "progress" set is used to asses progress over the years in the DARPA EARS project. For system development the 2002 evaluation data set (*eval02*) was used.

The audio data is parameterised using 13 PLP features augmented with their first and second order derivatives. Vocal Tract Length Normalisation (VTLN) was used in training and test by warping the filterbank. Cepstral mean and variance normalisation was applied. All acoustic models were built using discriminative training based on the Minimum Phone Error (MPE) criterion [2].

A detailed description of all system components is provided in [3]. A review of previous work on CU-HTK Conversational Telephone Speech can be found in [4].

3. SYSTEM COMBINATION IN LVCSR SYSTEMS

Most CTS LVCSR systems that are optimised for accuracy (rather than runtime speed) rely on system combination to achieve state-of-the-art performance. For example, all systems entered in the RT-03 CTS evaluation employed system combination.

System combination for ASR was introduced in [5] with the ROVER tool developed to combine the word-level output generated by independent LVCSR systems. Significant gains in accu-

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

In *Proc. ICASSP2004*, volume 1, pages 249–252.

© IEEE 2004

racy were achieved by combining the outputs of different participants in the evaluations. Starting with [6] participants employed hypothesis combination inside their systems.

ROVER operates by aligning the word sequences generated by the different systems based on a Levenshtein distance metric. The resulting word graph consists of a series for arc sets where each arc in a set represents a hypothesis from one of the systems (either a word or "-" representing a deletion). One of the words is selected either using simple voting or based on confidence scores assigned to the words by the different systems.

An extension of the basic ROVER approach is Confusion Network Combination (CNC) [8] which uses Confusion Networks [7] instead of 1-best hypotheses as input for the combination process. Each confusion set (or "sausage") contains the most likely competing word hypotheses from one system with associated posterior probabilities. Instead of aligning sequences of words as in ROVER, in CNC sequences of confusion sets are aligned. Based on the alignment the decision makes use of the full posterior distributions by summing the posteriors generated by the systems.

To maximise the effectiveness of system combination the systems to be combined should ideally have similar, low error rates but exhibit significantly different error patterns. Empirical evidence shows that the most effective way to produce such complementary systems is to built the models totally independently in separate research groups using different software and different training procedures. This has been demonstrated by NIST in post-evaluation experiments [5] where substantial improvements were achieved by combining the submissions from all the participants. Alas, this is not really feasible for normal systems.

4. HIGH-LEVEL SYSTEM STRUCTURE

The CU-HTK CTS systems consists of two main stages: lattice generation with adapted models and lattice rescoring in multiple branches. The aim of the lattice generation is twofold. Firstly, it provides large high-quality lattices that restrict the search space in the subsequent rescoring stage. Secondly, it provides supervision information for 1-best and lattice-based adaptation in each of the branches of the rescoring stage.

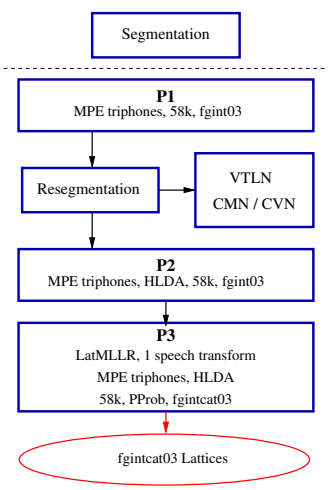

Fig. 1. Structure of the lattice generation stage

The audio data is segmented using a GMM-based procedure [9]. The actual lattice generation is performed with three full decoding passes. The first pass (P1) generates a transcription (using MPE HLDA trained triphones and the word 4-gram LM). This initial transcription is used to choose a VTLN warp factor for each conversation side. The second pass (P2) uses MPE VTLN HLDA triphones with the same LM to create small lattices for use in lattice MLLR [10] in the next pass. P3 uses the same models as P2 adapted using lattice MLLR (1 transform each for speech and silence). Large word lattices are generated with the word 4-gram LM interpolated with the class trigram.

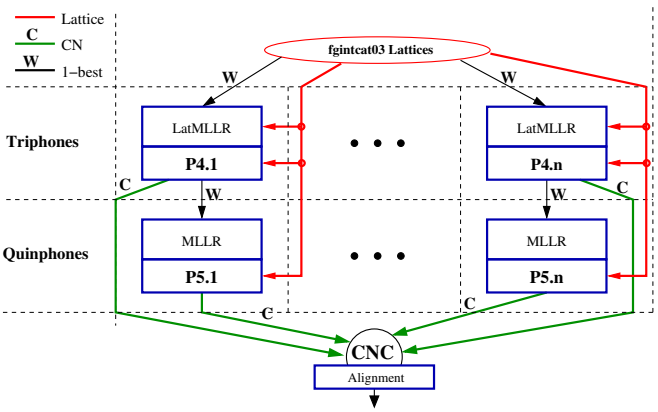

Fig. 2. Generic structure of lattice rescoring stage

The aim of the lattice rescoring stage is to generate hypotheses using a number of different acoustic models. For a schematic view of this stage see Figure 2. It consists of a number of branches each split into a triphone and a quinphone part. Each branch corresponds to a particular model building technique (e.g. SAT). All models are cross-word context-dependent (± 1 phone for the triphones, ± 2 for the quinphones) and the triphones are position-independent while the quinphones use word-boundary position information in the decision-tree state clustering. In each branch the triphones are adapted using a full-variance transform and up to 4 speech MLLR transforms. These transforms are estimated in a lattice-based framework that relies on model-marked lattices, which are generated for each branch using the respective triphones adapted with global 1-best MLLR. The 1-best supervision is taken from the output of the lattice generation stage.

The large rescoring lattices produced in the first stage are then rescored with the adapted triphones and new lattices are generated on output. These lattices are then converted into confusion networks [7] which provide a compact representation of the most likely alternative word hypotheses as input for the system combination. A 1-best minimum word error hypothesis is also extracted from the confusion networks for us in the quinphone rescoring part.

In the quinphone part only 1-best global MLLR and full-variance transforms are estimated. Again the adapted models are used to rescore the original rescoring lattices generating output lattices and then confusion networks.

All the confusion networks generated (2 from each branch) serve as input to the system combination based on CNC.

5. PERFORMANCE OF MODELLING TECHNIQUES

In this section the performance gains achieved by a number of advanced modelling techniques are illustrated in the context of a system based on the structure introduced above. All experiments were performed on the eval02 data set with automatic segmentations.

	purpose	WER
P1	supervision for VTLN	34.2
P2	supervision for MLLR	28.4
P3	lattice generation	24.8

Table 1. %WER on eval02 for lattice generation stage

Table 1 shows the error rates of the three passes of the lattice generation stage. The gain from performing VTLN is very substantial on this task (5.8% absolute). The difference between the accuracy of P3 and P2 (3.6% abs.) is the result of performing global MLLR adaptation, using pronunciation probabilities, smoothing the LM with the class trigram and wider pruning beams.

In the lattice rescoring stage a number of modelling techniques were employed. These were

SAT Speaker Adaptive training using constrained MLLR transforms estimated and fixed before MPE re-estimation (details in [11]).

HLDA Heteroscedastic Linear Discriminant Analysis. The PLP features were augmented with third order derivatives and projected down to 39 dimensions (details in [12]).

SPron Based on the multiple pronunciation dictionary and alignment of the training data a single pronunciation was selected for each word using a probabilistic model [13].

Based on these techniques triphone and quinphone models for four branches were trained using MPE on VTLN warped features:

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

The triphone models in branch B correspond to the set used in the lattice generation stage. The performance of each of the resulting 8 model sets before and after confusion network decoding is shown in Table 2

		Viterbi	+CN
P4.A	SAT tri	23.4	23.0
P4.B	HLDA tri	23.9	23.6
P4.C	SPron tri	23.4	23.4
P4.D	non-HLDA tri	25.7	24.8
P5.A	SAT quin	23.8	23.0
P5.B	HLDA quin	24.1	23.5
P5.C	SPron quin	23.9	23.3
P5.D	non-HLDA quin	26.0	24.6

Table 2. %WER on eval02 for individual models

The use of HLDA gives a significant gain (branch B vs. D) of 1.8% and 1.9% abs. for triphones and quinphones, respectively. However, HLDA reduces the benefit of CN-decoding (tri: 0.3/0.9; quin: 0.6/1.4), since the HLDA transform significantly changes the range of acoustic scores and thus affects the distribution of posteriors that are combined.

Both SAT and SPron consistently outperform the basic HLDA model set with improvements of 0.5% for triphones and about 0.3% for quinphones. The use of CN-decoding seems to be less effective if SPron dictionaries are used (particularly for P4). CN-decoding sums the posteriors of multiple pronunciations of a word, but in the SPron case the gain from this effect will be reduced.

In general the gains from CN decoding are smaller in the triphone pass, this is assumed to be a result of using lattice MLLR with multiple transforms which distorts the posterior distributions.

6. CHOOSING MODEL SETS FOR COMBINATION

For the integration in the full system it is not only the absolute performance of the individual models that is relevant but also the extent to which the outputs are complementary and thus effective in system combination. The results of pairwise system combination for each pair of triphone systems are given in Table 3. The best result is achieved by combining the two best single systems (SAT and SPron), closely followed by the combination of the best and the worst performing system (SAT and non-HLDA). It is surprising how effective the combination of the non-HLDA system with any of the three HLDA systems is despite the fact that its word error rate is 1.8% absolute below the SAT system's.

System (P4)	A SAT	B HLDA	C SPron	D non-HLDA
	23.0	23.6	23.4	24.8
+A		23.1	**22.6**	22.7
+B			22.9	23.3
+C				22.8

Table 3. %WER of individual triphone systems and pairwise combination on eval02 after lattice-MLLR/FV and CN

Instead of just combining pairs of systems, three systems can be used to improve performance further as shown in Table 4 which give results for all triphone 3-way combinations and the full 4-way combination for comparison. The best combination uses the SAT, SPron and non-HLDA branches (A+C+D).

Systems (P4)	WER
A+B+C	22.7
A+B+D	22.8
A+C+D	**22.4**
B+C+D	22.7
A+B+C+D	22.6

Table 4. %WER 3-way combination of triphone systems (eval02)

The inclusion of the quinphone model output results in a further performance improvement. The result of combining triphone and quinphone outputs is shown in Table 5 for the three branches most effective in combination at the triphone levels. The combination of all four branches, corresponding to 8-way combination, is given for comparison.

Based on the combination results shown in Table 5 it was decided to use the SAT, SPron and non-HLDA branches in the final 2003 CU-HTK systems leading to 6-way system combination. Figure 3 shows the resulting structure of the lattice rescoring stage.

Systems (P4 & P5)	WER
A+C+D (6-way)	21.7
A+B+C+D (8-way)	21.8

Table 5. Combination of triphone and quinphone systems (eval02)

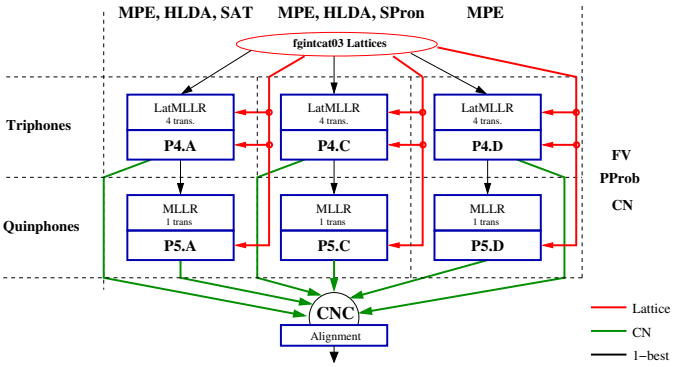

Fig. 3. 2003 system lattice rescoring stage

7. SYSTEM PERFORMANCE

The performance of the individual stages of the final 2003 system on the official eval03 set is given in Table 6. The eval03 set consists of Switchboard II phase 5 and Fisher data. The performance of the individual passes and the final results show very similar patterns to the results on the development set reported above.

		Sw2P5	Fisher	Total
P1	trans for VTLN	37.7	27.9	33.0
P2	trans for MLLR	31.8	22.6	27.4
P3	lat gen	27.5	19.3	23.5
P4.A	SAT tri	25.4	18.2	21.9
P4.C	SPron tri	25.6	18.5	22.2
P4.D	non-HLDA tri	27.4	19.6	23.7
P5.A	SAT quin	25.5	18.4	22.1
P5.C	SPron quin	25.7	18.7	22.3
P5.D	non-HLDA quin	27.5	19.6	23.7
CNC	P4.[ACD]+P5.[ACD]	24.1	17.1	20.7

Table 6. %WER on eval03 for all stages of 2003 system

The overall system ran in 187 times real time and the final confidence scores, which were estimated based on the confusion network posteriors, had a Normalised Cross Entropy (NCE) of 0.318.

8. CONCLUSIONS

In this paper the development of the 2003 CU-HTK LVCSR system for the transcription of Conversational Telephone Speech has been described. A system structure geared towards system combination and employing sophisticated adaptation was introduced.

The effectiveness of a number of modelling techniques (SAT, HLDA, SPron) was investigated and their performance compared in the framework of the full LVCSR system. Based on the single model performance and their contribution in system combination a subset of models was chosen to be integrated in the final 2003 CU-HTK system. This system has three branches (corresponding to SAT, non-HLDA and SPron), uses triphones and quinphones and generates the final word hypotheses via 6-way system combination. In the DARPA/NIST 2003 Rich Transcription evaluation this system exhibited state-of the-art performance, coming second on the eval03 set and first on the progress set with no statistically significant difference between the top two systems.

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 S.E. Tranter and K. Yu.

9. REFERENCES

[1] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-dependent Mixtures" in *Proc. HLT*, 2003.

[2] D. Povey and P.C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training" in *Proc. ICASSP*, 2002.

[3] P.C. Woodland et al., "CU-HTK English CTS Systems" in *Proc. Rich Transcription Workshop*, 2003.

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

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

[6] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, E.W.D. Whittaker and S.J. Young, "The 1997 HTK Broadcast News Transcription System" in *Proc. DARPA Broadcast News Transcription Workshop*, 1998.

[7] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization" in *Proc. Eurospeech*, 1999.

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

[9] S.E. Tranter, K. Yu, G. Evermann, and P.C. Woodland, "Generating and Evaluating Segmentations for Automatic Speech Recognition of Conversational Telephone Speech" *Proc. ICASSP*, 2004.

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

[11] L. Wang and P.C. Woodland, "Discriminative Adaptive Training using the MPE Criterion" in *Proc. ASRU*, 2003.

[12] X. Liu, M.J.F. Gales, and P.C. Woodland, "Automatic Complexity Control for HLDA Systems" in *Proc. ICASSP*, 2003.

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