THE 1998 HTK SYSTEM FOR TRANSCRIPTION OF CONVERSATIONAL TELEPHONE SPEECH

T. Hain, T.R. Niesler , E.W.D. Whittaker and P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
{th223, trn, ewdw2, pcw}@eng.cam.ac.uk

ABSTRACT

This paper describes the 1998 HTK large vocabulary speech recognition system for conversational telephone speech as used in the NIST 1998 Hub5E evaluation. Front-end and language modelling experiments conducted using various training and test sets from both the Switchboard and Callhome English corpora are presented. Our complete system uses reduced bandwidth analysis, side-based cepstral feature normalisation, vocal tract length normalisation (VTLN), triphone and quinphone hidden Markov models (HMMs) built using speaker adaptive training (SAT), maximum likelihood linear regression (MLLR) speaker adaptation and a confidence-score based system combination. A detailed description of the complete system together with experimental results for each stage of our multi-pass decoding scheme is presented. The word error rate obtained is almost 10% absolute better than our 1997 system on the development set.

1. INTRODUCTION

Transcription of conversational telephone speech is a complex task, which has to deal with many severe degradations in speech quality. These degradations continue to lead to word error rates in the range of 30 to 50 %, which are almost twice as high as for other difficult tasks like Broadcast News Transcription [8]. The problems result from limited bandwidth, distorted audio channels, cross-talk and other acoustic interference, as well as highly variable speaking rates and conversational styles in which grammatical rules are less important.

Current experiments for conversational telephone speech are usually conducted on three corpora distributed by the Linguistic Data Consortium (LDC) : Switchboard-I (Swbd-I), Switchboard-II (Swbd-II) and Callhome (CHE). Both Switchboard corpora consist of telephone conversations within the USA between mutual strangers. For Swbd-I speakers are given a topic, whereas for Swbd-II the topic is merely suggested. CHE data consists of calls to friends or relatives abroad. This leads not only to severe acoustic channel distortions caused by long distance telephone connections, but also to a higher number of non-English (and hence unknown) words. Furthermore, multiple speakers per conversation side are not uncommon. These factors usually lead to an at least 10% difference in word error rate between Switchboard and Callhome recognition tests.

The Swbd-I, Swbd-II and CHE corpora are the subject of the yearly Hub5 evaluation conducted by the National Institute for Standards and Technology (NIST). In the following sections we describe the system prepared for participation in the 1998 Hub5E evaluation, and present its subsequent performance.

The remainder of this paper is organised as follows: First we give a brief overview over our system and the development objectives. Then we describe our frontend experiments, including analysis bandwidth, cepstral feature normalisation and vocal tract length normalisation (VTLN). Subsequently we present the speaker adaptation tests using the new frontend, as well as language modelling experiments. The final section details the overall system performance.

2. SYSTEM OVERVIEW

The basis for our 1998 development procedure was formed by our 1997 conversational telephone speech transcription system [7]. This system employed gender independent decision-tree state-clustered triphone models, a 3-gram language model trained on 2MW from Swbd-I and CHE, and a 22K word dictionary based on the LIMSI 1993 WSJ pronunciation dictionary.

Standard techniques for telephone speech have been employed, with the only particular refinement being the introduction of VTLN. All available bandwidth has been used with per segment cepstral mean normalisation.

During the development of our 1997 system we found the vocal tract length normalisation (VTLN) process to give unreliable results in terms of word error rate (WER) across speakers and test sets. In particular, the performance gain when using VTLN was considerably lower than expected especially for the 1997 Hub5E evaluation set. Furthermore, frontend processing of telephone speech data that does not take into account the usually very short speech segment durations and special characteristics of the transmission channel is clearly suboptimal. These issues have been addressed in a series of experiments, and subsequent improvements have been implemented and tested in our current system.

Our 1998 system uses eight-pass decoding with multiple gender independent and gender dependent state-clustered triphone and quinphone HMM model sets, and multiple stages of speaker adaptation.

Each frame of input speech is represented by a 39 dimensional feature vector that consists of 13 (including c_0) MF-PLP cepstral parameters and their first and second differentials. Experiments described in section 3.1 suggested the use of reduced bandwidth analysis and cepstral mean and variance normalisation per conversation side.

Three different types of HMM model sets are used. First, a gender independent state clustered triphone model set has been built and trained using a subset of Swbd-I containing 65 hours of speech (WS96train). The resulting system contains 6039 speech states with up to 12 Gaussian mixture components each. The final model set (M1) was obtained from this by further reestimation and mixture splitting steps using a 180-hour training set (h5train98) consisting of 163 hours of speech from Swbd-I and 17 hours from CHE. It was found that 16 mixture components per speech state was optimal. The M1 model set has been used in the first decoding pass to obtain transcripts for gender detection and VTLN warp factor computation.

Second, a gender independent triphone model set using VTLN warped training data has been obtained in a similar fashion to M1. Gender dependent versions have then been derived by a gender dependent reestimation step where the silence and short pause models were kept unchanged. This model-set pair, further referred to as M2, was used in the second and the third recognition passes.

Finally, decoding passes 4-7 used a gender dependent pair of quinphone VTLN-trained HMM models (M3). These models were trained using h5train98 with a further speaker adaptive training (SAT) [6] stage. The resulting set contained 8763 speech states, each characterised by a 16 component mixture Gaussian.

In passes 3-7, maximum likelihood linear regression (MLLR) [3] was employed for updating both means and variances per speaker. Whereas one global MLLR transform was used per conversation side in passes 3 and 4, subsequent stages used a maximum of 2, 4 and 8 transforms per side.

The final stage combines two different system outputs according to a computed confidence score for each word. The confidence scores were generated using an N-best homogeneity measure found using the 1000-best hypotheses from the lattices generated at the appropriate stage. A decision tree pruned using 10-fold cross-validation was used to convert the N-best homogeneity scores to confidence probabilities. This decision tree has been trained on the development data also using 10-fold cross validation. System output of the best triphone system (pass 3) and the best quinphone system (pass 7) are combined using ROVER [2].

3. FRONT-END EXPERIMENTS

For fast turnaround on front-end experiments, a small subset of the Swbd-I corpus was chosen for training. This subset (referred to as MiniTrain) covers 398 sides containing 17.8 hours of speech and is approximately gender balanced. For testing a gender balanced half-hour set (MTtest) containing Swbd-I data set was chosen. All front-end experiments have been conducted using a 2MW Switchboard trigram language model. The following sections cover experiments on analysis bandwidth, cepstral normalisation, and VTLN.

3.1. Coding Bandwidth and Cepstral Normalisation

Due to the special characteristics of telephone channels, the lower and upper frequency regions are either distorted or blocked by filtering operations. We compared systems using Mel-scale Filterbanks within the full 4kHz region and a reduced region between 125-3800 Hz. This not only has the effect of masking out potentially irrelevant portions of the spectrum, but also different filter bank bandwidths and centres.

Since the individual filter banks get very narrow in terms of spectral samples (especially together with vocal tract length normalisation as described in section 3.2), the FFT resolution has been doubled using zero padding. Speech recognition systems designed for read or even Broadcast News data usually apply a per segment cepstral mean normalisation scheme to reduce the effects of constant channel characteristics. However, for telephone conversations the average utterance duration is less than 3 seconds, thus giving poor estimates for segment means. To overcome this, the mean has been calculated over a complete conversation side.

Results in table 1 show the performance of HMM model sets trained on MiniTrain for different bandwidths using both cepstral mean subtraction strategies. Surprisingly the reduced bandwidth system performs worse on the MTtest set. Nevertheless both coding strategies show a gain of about 1%, and this is slightly higher with the reduced bandwidth.

	0-4000Hz	125-3800Hz
Seg-CMN GI	46.58	47.33
Side-CMN GI	45.67	46.17

Table 1: % Word Error Rates (WER) for full and reduced bandwidth coding using models trained on the MiniTrain set and tested on MTtest

We further tested the performance of variance normalisation in conjunction with side-based mean normalisation using several different techniques. A standard segment based scheme has been compared with side-based variance normalisation and normalisation using a time constant decay. Each feature vector component was normalised to obtain a target variance, which has been selected to be the overall test data variance. A linear and a non-linear scheme masking out feature scaling factors smaller than one have been tested. Linear side-based variance normalisation produced the best results.

	0-4000Hz	125-3800Hz
Side-CMN, Side-CVN , GI	44.82	44.35
Side-CMN, Side-CVN , GD	44.33	43.00

Table 2: % WER on MTtest using different bandwidth and gender independent (GI) and gender dependent (GD) MiniTrain model sets.

Table 2 shows the effect of side based variance normalisation using both full and reduced bandwidth coding on gender dependent and gender independent models. The performance gain is high especially for the reduced bandwidth case. Reduced bandwidth coding outperforms full bandwidth analysis further using gender dependent HMM models.

3.2. Maximum Likelihood Vocal Tract Length Normalisation

Maximum likelihood vocal tract length normalisation [1] implements a linear frequency scaling of the speaker spectrum. The scale factor is obtained using a search procedure and is then applied in speaker specific feature stream computation. Speaker spectrum scaling can also be implemented by scaling the Mel filterbank centre frequencies with the inverse warping factor. Smoothing of the upper frequency filterbank contents is required using scale factors larger than one. Instead of achieving this by mirroring the

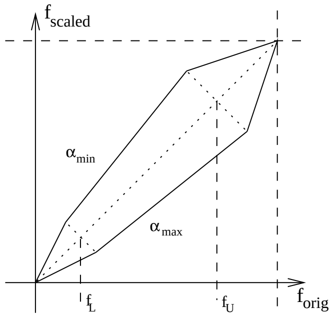

Figure 1: Piecewise linear VTLN frequency scaling function for warp factors α. f_L and f_U denote lower and upper threshold frequencies.

contents of the upper frequency contents, our new approach introduces a piecewise linear warping function with lower and upper cut-off frequencies (see figure 1). The upper threshold has improved the stability of our implementation (table 3), while the lower cut-off frequency affects performance only slightly. Warp factors are found by conducting a parabolic search over data likelihoods versus warp factors. Given a previously obtained word level transcript, the average per-frame log-likelihood given some HMM model set is computed by feature recomputation and alignment with the transcripts. The next warping factor is then selected and the procedure repeated. Since the per-frame log-likelihood tends to be a parabolic function of the warp factor, a suitable search method has been chosen to allow the estimation of the warp factor in less than two times realtime.

	test	train & test
old	43.21	42.66
new	42.52	41.56

Table 3: % WER comparison for different VTLN implementations on MTtest using full bandwidth coding. Test denotes test-set VTLN only, train & test denotes single iteration VTLN models

Warp factors are computed using standard HMM models trained on a particular dataset for the use of VTLN in training. New models are generated by single pass retraining with the appropriately warped training data. Since multiple iterations of VTLN training are necessary to allow the warp factor distribution to converge, models generated in one iteration serve as a warp factor estimator as well as the base for single pass retraining in the next iteration.

Experiments on both bandwidths are shown in table 4. In the full bandwidth case the warp factor distribution already settled after two iterations, whereas four iterations have been necessary on reduced bandwidth models. Even though in the gender independent case full-band coding seems to perform equally well as reduced-band coding, the gain on gender dependent modelling for reduced-band coding still is 0.4% higher. Since the relative improvement is small, this result has been cross-checked with another test set.

		0-4000Hz	125-3800Hz
GI	test	43.21	43.18
GD	test	41.92	41.33
GI	train/test	41.45	41.61
GD	train/test	40.75	40.20

Table 4: % WER for systems trained on MiniTrain and tested on MTtest. VTLN warping in training and test both for GI and GD models for full and reduced bandwidth coding.

4. SPEAKER ADAPTATION

The more robust implementation of VTLN together with gender dependent modelling reduced the improvement in word error rate achieved with MLLR speaker adaptation significantly. Our 1997 system achieved a gain of 4.8% on the 1997 Hub5E evaluation set using only global mean and variance speech transforms, and a further 1.6% by subsequent iterations with larger numbers of speech transforms. In comparison, only 2.5% improvement has been obtained using our 1998 front-end. The contribution of variance adaptation was only 0.2%, and this small figure may be attributed to to the effectiveness per-speaker variance normalisation. Subsequent MLLR iterations gave approximately similar improvements as for our 1997 system.

5. LANGUAGE MODELLING

Approximately 3 million words of Switchboard and Call Home English transcriptions were available for language model training (h5trainLM). From this, a 27k word recognition vocabulary containing only English words was determined. Furthermore, backoff bigram (bgH5), trigram (tgH5) and 4-gram (fgH5) models were trained from h5trainLM. These models contained 426k bigrams, 292k trigrams and 281k 4-grams. To evaluate the effect of the increase in training data, a trigram tgH5_97 was built using the approximately 2 million words of Switchboard transcriptions available in 1997.

Using the 27k wordlist, bigram (bgBN), trigram (tgBN) and 4-gram (fgBN) models were trained from Broadcast News data ranging in epoch from January 1992 to December 1997. These models contained 3.8 million bigrams, 5.8 million trigrams and 6.4 million 4-grams.

Corresponding H5 and Broadcast News models were merged by linear interpolation into a single resultant language model file, allowing them to be used directly in the recognition search. Thus bgH5 was merged with bgBN to form bgint98, tgH5 with tgBN to form tgint98, and fgH5 with fgBN to yield fgint98.

Finally, a class-based trigram language model (cat98) was produced using 350 automatically generated word classes based on word bigram statistics [4]. Bigrams and trigrams are only added if they improve the training set leave-one-out perplexity [5]. Both the categories as well as the trigram category model were built using only h4trainLM. The model contained 75k bigrams and 231k trigrams. An optimal interpolation (in terms of eval97 perplexity) was produced between fgH5, fgBN and cat98 with respective weights of 0.42, 0.28 and 0.30 , and will be referred to as fgintcat98.

Table 5 displays the performance of these language models. Note that the 1997 NIST scoring conventions have been employed in WER calculation. The WER results for tgint98, fgint98, and fgintcat98 were obtained by rescoring latticed produced with tgint98.

LM	PP	WER
tgH5_97	98.3	-
tgH5	94.1	-
cat98	101.8	-
bgint98	101.7	45.8
tgint98	82.0	42.7
fgint98	79.2	42.3
fgintcat98	76.4	41.5

Table 5: Perplexity (PP) on eval97 and WER on eval97sub for various language models.

6. SYSTEM RESULT ANALYSIS

Table 6 shows the performance of the individual stages on a subset of the 1997 Hub5E evaluation set (eval97sub) and the full 1998 Hub5E evaluation set (eval98). The eval97sub set has been used for system development and consists of 20 conversation sides, equally divided into a Swbd-II and a CHE part. This set has been selected to give approximately equal performance as the full 1997 evaluation set. The eval98 set is gender balanced on Swbd-II data, but only contains 6 male speakers from CHE.

PASSES	Total	Swbd-II	CHE
P1	51.1	43.6	58.7
P2	44.6	36.5	52.8
P3	39.5	31.1	48.0
P4	38.1	29.9	46.4
P5	37.5	29.0	46.0
P6	37.3	29.1	45.6
P7	37.1	28.7	45.5
P8	36.6	28.5	44.7

(a)

PASSES	Total	Swbd-II	CHE
P1	49.3	47.0	51.6
P2	45.6	42.9	48.2
P3	42.6	39.9	45.3
P4	40.9	38.3	43.4
P5	40.5	37.9	43.2
P6	40.4	37.7	43.0
P7	40.3	37.7	42.8
P8	39.5	36.7	42.2

(b)

Table 6: % WER for the eval97sub set (a) and the eval98 (b) set for each decoding pass P1-P8. Word error rates are computed using the 1998 Hub5E scoring rules.

In the first pass (P1) a word level transcript has been obtained using M1 models and the tgint98 language model. In the second pass (P2) all VTLN warping factors for all sides are computed using gender dependent warp estimation models and the output of the first pass. Secondly the likelihood for the best warp factor for both genders are compared and gender is selected according to the more likely model set. Whereas on eval97sub this gave no gender detection errors, this has not been the case on eval98, where 3 sides out of 80 have been misclassified.

Afterwards M2 models and the tgint98 language model are used to produce better MLLR supervision for the next stage. On the eval97sub gender dependent modelling plus VTLN brought a 6.5% gain in WER compared to only 3.7% on eval98.

In the third pass (P3) M2 models, MLLR speaker adaptation and the interpolated bigram model bgint98 are used to produce lattices, which are rescored and pruned using tgint98 and fgintcat98 language models. The gain of MLLR plus 4-gram language modelling was 5.1% on eval97sub compared to only 3% on eval98. The use of the M3 quinphone models with further MLLR passes brought 2.4% on eval97sub and 2.3% on eval98. Final system combination performed approximately equal on both sets with 0.8% in eval98 and only 0.6% on eval97sub.

The difference in performance is not easy to explain. Error rates on Switchboard in eval98 are unexpectedly high, whereas the CHE portion seemed to be less difficult than usual.

7. CONCLUSIONS

We have shown a significant improvement in performance based on more appropriate acoustic and language modelling concepts in our 1998 conversational telephone speech transcription system. To achieve this we added different bandwidth analysis, side-based cepstral feature normalisation, improved VTLN and SAT trained quinphone models. On the language modelling side a step to 4-grams and 3-fold interpolation was made. Huge variations in word error rate for different test sets however make it difficult to predict performance on unknown data.

8. ACKNOWLEDGEMENTS

We would like to thank BBN for providing the MiniTrain training and MTtest test set definitions. This project was sponsored by GCHQ.

9. REFERENCES

[1] Eide E. & Gish H. (1996). A Parametric Approach to Vocal Tract Length Normalisation. *Proc. ICASSP'96*, Vol. 1, pp. 346-348, Atlanta.

[2] Fiscus, J.G. (1997) A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). *Proc. IEEE ASRU Workshop*, pp. 347-352, Santa Barbara.

[3] Gales M.J.F. & Woodland P.C. (1996). Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249-264.

[4] Kneser R., & Ney H. (1993). Improved Clustering Techniques for Class-Based Statistical Language Modelling. *Proc. Eurospeech'93*, pp. 973-976, Berlin.

[5] Niesler T.R., Whittaker E.W.D & Woodland P.C. (1998). Comparison of Part-Of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition. *Proc. ICASSP'98*, pp. 177-180, Seattle.

[6] Pye D. & Woodland P.C. (1997). Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition. *Proc. ICASSP'97* , pp. 1047- 1050, Munich.

[7] Woodland P.C., S. Kapadia, H.J. Nock & S.J. Young (1997). The HTK System for the March 1997 Hub 5 Evaluations. Presented at the *DARPA Hub5E Conversational Speech Recognition Workshop*, May 13-15,1997, Baltimore, Maryland.

[8] Woodland P.C., Hain T. , Johnson S.E., Niesler T.R., Tuerk A.,Whittaker E.W.D. & Young S.J. (1998) The 1997 HTK Broadcast News Transcription System. *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 41-48, Lansdowne, Virginia