ISCA Archive
http://www.isca-speech.org/archive

ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

Use of Model Transformations for Distributed Speech Recognition

Naveen Srinivasamurthy, Shrikanth Narayanan and Antonio Ortega

Integrated Media Systems Center, Dept of EE-Systems
University of Southern California
Los Angeles, CA 90089-2654
[snaveen,shri,ortega]@sipi.usc.edu

Abstract

Due to bandwidth limitations, the speech recognizer in distributed speech recognition (DSR) applications has to use encoded speech – either traditional speech encoding or speech encoding optimized for recognition. The penalty incurred in reducing the bitrate is degradation in speech recognition performance. The diversity of the applications using DSR implies that a variety of speech encoders can be used to compress speech. By treating the encoder variability as a mismatch we propose using model transformation to reduce the speech recognition performance degradation. The advantage of using model transformation is that only a single model set needs to be trained at the server, which can be adapted on the fly to the input speech data. We were able to reduce the word error rate by 61.9 %, 63.3 % and 56.3 % for MELP, GSM and MFCC-encoded data, respectively, by using MAP adaptation, which shows the generality of our proposed scheme.

1. Introduction

The recent explosion in mobile computing and communication devices has generated a wide interest in the distributed speech recognition paradigm. The more straightforward method (sometimes called network speech recognition (NSR)) shown in Figure 1 involves encoding speech using a "standard" speech encoder before transmission to the client. Speech compression algorithms are, however, typically designed to effect minimal degradation in the *perceived* quality of the decoded speech. Although the signal distortions introduced by these speech compression techniques may be perceptually irrelevant, they may be detrimental in the context of ASR. Ideally, since compressed speech is used for recognition (classification) at the server, the compression technique should be optimized to introduce minimal degradation of speech recognition accuracy. This can be accomplished by adopting a "client-server" system, where speech features are extracted at the client (device), then compressed and transmitted to a remote server hosting the speech recognizer as shown in Figure 2. Recognition from the encoded feature vectors performs better than when encoded speech is used for recognition. However there is some recognition degradation when compared to clean (uncompressed) feature vectors.

The effect of compression of speech (features) for speech recognition has been reported previously [1, 2, 3, 4]. In previous work [3], we developed simple encoders to compress the mel frequency cepstral coefficients (MFCCs) derived from the speech utterance. These experiments assumed a scenario wherein the models, trained using unquantized features, were kept fixed. For the TI-46 digit database, the baseline word error rate of 0.24 % for uncompressed speech degraded by 447 % (1.15 % error) for a MELP coder at 2.4 kbps and by 500 % (1.26 % error) at 1 kbps when the MFCC features were directly quantized. In this work we wish to investigate the complementary problem of optimizing the speech recognizer to take into account that it is operating on compressed speech. The novelty of our approach is that we view the differences in the compression underlying the train and test utterances as a mismatch which results in degradation of the classifier's performance. This mismatch can be reduced by using robust adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) [5] or Bayesian adaptive techniques [6] to modify the reference models using the observed quantized MFCCs.

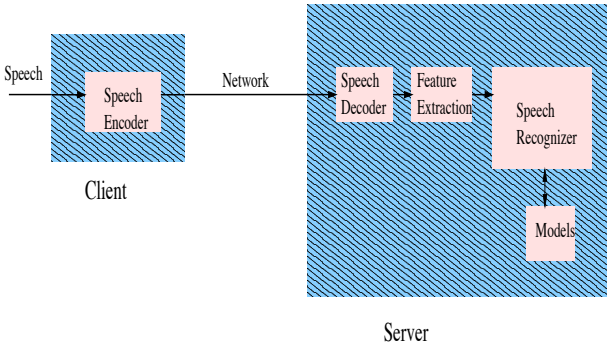

Figure 1: Distributed speech recognition setup with a standard speech encoder.

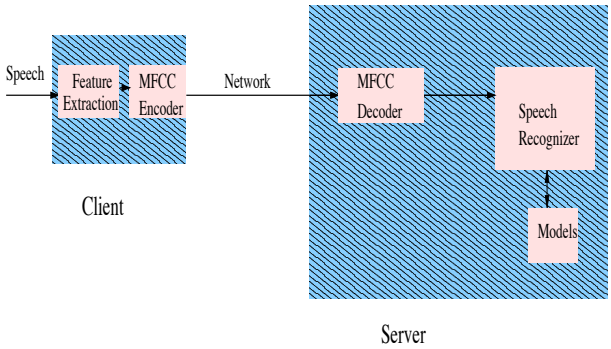

Figure 2: Distributed speech recognition setup with a MFCC encoder.

For the TIDIGITS database we achieved a reduction of 56.3% in word error rate(WER) by adapting the clean models

This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152.

to the MFCC encoded data. Improvements of 61.9% and 63.3% were achieved when the speech was compressed by MELP and GSM. Section 2 contains the details of the adaptation schemes. Experimental details are presented in Section 3. Section 4 presents our results and our conclusions are presented in Section 5.

2. Model Adaptation

One of the major problems for robust speech recognition is the mismatch between the training and testing conditions. Speech recognition performance, with speech models trained on clean data, significantly degrades when the test utterances are noisy (channel noise, ambient environment noise). Similarly the performance is also degraded due to long term and short term speaker variations. It is well known that speaker dependent models usually outperform speaker independent models. To improve robustness, techniques proposed involve (i) finding invariant features; (ii) allowing model parameters/feature vectors to vary within a neighborhood specified by the training data; (iii) transforming the models so they are more likely to have produced the observed data; (iv) incorporate newly acquired application specific data into existing models.

With wider use of speech recognition applications, especially in mobile devices, we have an additional source for mismatch, namely speech encoding. The distortion introduced by speech encoders can also be thought of as a mismatch between the training and testing conditions. It is relatively easy to remove this mismatch by the use of a family of models each trained with data from different encoding schemes, and choose the one that best matches the unknown test data. However, such schemes are not attractive since it might not be possible to have models trained for all different compression schemes because the choice of the compression scheme used by the client may be made dynamically depending on the computational resources/load at the client and the quality of service (QoS) it wishes to provide the user. Scalable encoders, which could be combined with scalable recognition schemes [7], wherein the recognition is refined in every pass with more data (and/or better models) until a satisfactory decision (say in the likelihood sense) can be made, further complicates the creation of predefined models. Depending on the optimization criteria used for compression (recognition performance or human perception), more variability in the compression schemes used by the different clients can be expected.

This mismatch introduced by the choice of different speech compression schemes can be solved in similar manner as other mismatches. The models at the server can be trained using clean speech (or a particular compression scheme) and we can alleviate the mismatch between testing and training phases by the use of model transformation/adaptation to optimize classification by ensuring that the transformed/adapted models are more likely to have produced the observed data. Note that simple signal processing techniques are not likely to be helpful as the distortion introduced by compression is not invertible. However adaptation schemes, which operate on the models rather than the input speech are more likely to be able to reduce the mismatch.

The two popular adaptation techniques which have been used previously are MLLR and Maximum a posteriori (MAP) estimation. In the MLLR technique a transformation is computed for the means and variances of the different mixture components after observing the new data. Regression classes are defined to facilitate transformation even when a limited amount of data is observed. MAP in contrast assumes that model parameters are random and have a prior distribution. The observed data can be combined with the existing models to obtain new models by maximizing the posterior density of the models given the observed data. Unlike MLLR, in MAP we can modify not only the means and variances of the Gaussian mixtures but can also modify the mixture weights, the initial probabilities and the transition probabilities. For both methods, adaptation can be carried out either in batch mode or in an incremental manner. In batch mode adaptation (or supervised adaptation) the transcription corresponding to the unknown utterance is available. Incremental adaptation (or unsupervised adaptation) does not require the transcription and the result of recognition is used as the "true" transcription of the unknown utterance.

3. Experimental Setup

The experiments were carried out on the TIDIGITS corpus using HTK 3.0 speech recognizer, with MFCCs as the front end. The database consists of variable length connected digit utterances (1 to 7 digits per utterance). The "train" part of the database consisted of 8623 utterances spoken by 55 male and 57 female speakers and the "test" part of the database consisted of 8700 utterances spoken by 56 male and 57 female speakers (the train and test speakers were different). Context-independent digit models were initially trained (on the server) using clean speech from the "train" part of the database. A silence model was used before and after the digit utterance to take care of the pre and post utterance silence. In addition a short pause model was used to account for inter-digit short pauses. The testing (using utterances from the "test" part of the database) was carried out using MELP compressed speech, GSM compressed speech and the MFCC encoder proposed in [3]. The MFCC encoder was used at two different rates 2.07 kbps (denoted MFCC-HR) and 1.22 kbps (denoted MFCC-LR). The baseline performance was determined by using "matched" models for the different compression schemes, i.e., the training was done using speech encoded by the same method as that used during the testing phase. The original database contains speech sampled at 20 kHz, however both MELP and GSM require the input speech to be sampled at 8 kHz. One method to overcome this would be downsample the speech to 8 kHz, encode the speech and then upsample the decoded speech back to 20 kHz, however when this method was used the performance obtained was poor. The reason for this could be that the spectrum of the upsampled speech is flat from 4 kHz to 10 kHz while the spectrum of the original speech was not. To overcome this we can downsample all the speech (training and testing) to 8 kHz and perform all our experiments using this downsampled data. Now the training phase also uses downsampled speech to build the initial models. For consistency the MFCC encoder also was used with the downsampled speech data. The experiments were carried out for two different settings (i) unsupervised MLLR adaptation and (ii) supervised MAP adaptation. For the unsupervised MLLR adaptation, the models were adapted once every 20 utterances. For the supervised MAP adaptation the original models were adapted for each speaker individually, i.e., part of the testing data from each speaker was used to adapt the original models to that particular speaker and the adapted models were used to recognize the test utterances of that speaker. The results for the different experiments are shown in Tables 1 to 4 for the different compression schemes, and for the baseline recognition on uncompressed speech.

Compression	Clean Models (E_{CM})	Clean Models + MLLR (E^A_{CM})	Matched Models (E_{MM})	Matched Models + MLLR (E^A_{MM})	MLLR gain
Clean speech	1.88 (7.56)	1.57 (6.57)	-	-	16.5%
MELP	3.14 (12.07)	2.32 (8.70)	2.70 (10.47)	1.87 (8.53)	26.1%
GSM	2.50 (8.76)	1.73 (7.33)	2.29 (8.61)	1.55 (6.91)	30.8%
MFCC-LR	4.81 (14.78)	2.24 (8.49)	2.70 (10.25)	1.85 (8.08)	53.4%
MFCC-HR	2.10 (8.06)	1.60 (6.82)	2.05 (7.87)	1.58 (6.87)	23.8%

Table 1: Word error rate (in percentage) for supervised MLLR adaptation. String error rate (in percentage) is shown in brackets. The improvements in MLLR are decrease (in percentage) in word error rate with respect to clean model results.

4. Results and Discussion

4.1. Adaptation with a Single Model

Table 1 shows the results for the unsupervised MLLR adaptation experiment. We observe that consistently for all the compression schemes MLLR adaptation results in good improvements in the recognition performance. The results after adaptation are in fact better than when "matched" models are used. The is because we are using unsupervised adaptation and updating the models once every 20 utterances, and the utterances from each speaker are together, so we are benefiting from inter utterance similarities (as indicated by the improved performance with adaptation on clean speech). To ensure that the comparisons are consistent we performed adaptation on the matched models (shown in column 5).

To show the advantage of adaptation with a single model, we can compute the degradation before and after adaptation. These can be evaluated by comparing the recognition performance with clean models to the recognition performance with matched models (baseline, no mismatch in training and testing). The single model degradation before adaptation is defined as $D = (E_{CM} - E_{MM})/E_{MM} * 100$, where E_{CM} is the error when compressed data is used for testing and clean speech is used for training, E_{MM} is the error when compressed data is used for training and testing. Similarly the single model degradation after adaptation is defined as $D^A = (E^A_{CM} - E^A_{MM})/E_M M^A) * 100$ (E^A_{CM} and E^A_{MM} are defined as above except that adaptation is used). These degradations are shown in Table 2. Observe that there is significant degradation before adaptation for MFCC-LR. However after adaptation the degradation is reduced substantially. For MFCC-HR, by adaptation from clean models we get almost same performance as adaptation from matched models (1.60 % vs. 1.58 %). These results imply that with adaptation from a single model we are not only able to reduce the absolute error rates but we are also able to reduce the degradation from matched conditions (for MELP and GSM the relative degradation increased but the absolute error rate decreased; for GSM the relative increase was very small). This result is very significant because it demonstrates that we do not need encoder specific models to be trained at the server, instead we can achieve the same performance with adaptation of models trained from clean speech.

The results of the supervised MAP adaptation are shown in Table 3. The supervised MAP results are better than the unsupervised MLLR results as expected. For MELP and GSM, MAP adaptation provides better results when compared to MLLR, however for the MFCC encoders the MAP performance does not provide as significant a decrease as for MELP and GSM (in fact for MFCC-LR, the MAP performance was worse than the MLLR performance). The reason for this could be that while MLLR does not model the initial parameters as a random vector MAP explicitly does. The MFCC encoder quantizes the

Compression	Degradation before Adaptation (D)	Degradation after Adaptation (D^A)
MELP	16.30	24.06
GSM	9.17	11.61
MFCC-LR	78.15	21.08
MFCC-HR	2.38	1.27

Table 2: Degradation (in percentage) in word error rate before and after adaptation for the different coding schemes. The degradation is with respect to using matched models for each compression scheme.

MFCCs directly and this means that the actual distribution of the encoded MFCCs is not a continuous distribution anymore but a discrete distribution. However in the MAP formulation the MFCCs are modeled as continuous distributions and the conjugate distribution which lies in the same class as the original distribution is used as the prior distribution. Therefore the MAP formulation is no longer optimal and this could be leading to the fact that we get less improvement with MAP than with MLLR for the MFCC encoders. Nevertheless, the improvement by using MAP adaptation is obvious from the results; we get more than 60% reduction using MAP for GSM and MELP. The reductions for the other methods are also significant.

Compression	Clean Models	MAP	MAP gain
Clean speech	1.86 (7.54)	0.67 (3.85)	64.0%
MELP	3.12 (12.05)	1.19 (6.06)	61.9%
GSM	2.48 (8.72)	0.91 (4.09)	63.3%
MFCC-LR	4.78 (14.73)	3.34 (10.89)	30.1%
MFCC-HR	2.08 (8.01)	0.91 (4.36)	56.3%

Table 3: Word error rate (in percentage) for supervised MAP adaptation. String error rate (in percentage) is shown in brackets. The improvements for MAP is decrease (in percentage) in word error rate with respect to clean model results.

4.2. Encoder Optimized for Recognition

As mentioned before, compression introduces degradation in recognition performance. The compression degradation can be found by comparing the results with compression to those obtained with clean speech. The compression degradation before adaptation can be found as $D_C = (E_{CM} - E_{UC})/E_{UC} * 100$ and the compression degradation after adaptation can be found as $D^A_C = (E^A_{CM} - E^A_{UC})/E^A_{UC} * 100$, where E_{UC} is the error when clean speech is used for training and testing (E^A_{UC} corresponds to the case when adaptation is used). These degradations are shown in Table 4 along with the rate required for the different compression schemes. The rate required by the MFCC encoders is significantly less than that required by GSM and is less

Compression	Degradation before Adaptation (D_C)	Degradation after Adaptation (D_C^A)	Rate (kbps)
MELP	67.02	47.77	2.4
GSM	32.98	10.19	13
MFCC-LR	155.85	42.68	1.22
MFCC-HR	11.70	1.91	2.07

Table 4: Degradation (in percentage) in word error rate before and after adaptation for the different coding schemes. The degradation is with respect to using clean speech for testing. Uncompressed speech requires 128 kbps and uncompressed MFCCs require 38.4 kbps.

than that required by MELP. However minimum degradation is introduced by the MFCC-HR encoder among all the compression schemes (WER only degraded from 1.57 % to 1.60 %). Also notice that after adaptation MFCC-LR encoder operating at half the rate of MELP actually provides better results than MELP. This justifies our initial claim that compression schemes optimized for recognition should be used to compress speech used with recognizers for better performance. Another important point to be noticed from this table is that consistently for all the encoding schemes the degradation after adaptation is lesser then the degradation before adaptation, which implies that adaptation is compensating for the compression mismatch in addition to compensating for other mismatches.

4.3. Effect of Adaptation Data

It is also important to find the dependency of the adaptation schemes on the amount of input data required. To find this we used MLLR adaptation in supervised mode and changed the amount of the adaptation data used. The experiments were carried out for clean data, MELP, GSM and MFCC-encoded data. The number of speakers in the test corpus was 113. The number of utterances chosen per speaker for adaptation was $1, 2, 4$ and 8 resulting in $113, 226, 452$ and 904 utterances used as adaptation data for the four different cases. The results of string error rate and word error rate are shown in Figure 3. Observe that with increased adaptation data the error rates decrease for the different encoders. Using more than 904 utterances provided no further improvement in performance. One of the drawbacks of this scheme is that improvements are seen only after sufficient adaptation data has been observed, which may not be practical in some situations. This is basically a problem of the adaptation schemes (MLLR and MAP) which we have used here. To overcome this it may be necessary to combine MLLR/MAP adaptation with other rapid adaptation schemes which can operate with lesser adaptation data.

5. Conclusions

In this paper we investigated the use of model adaptation to reduce the speech degradation introduced by encoding speech before recognition. We showed that we were able to improve the recognition performance for MELP, GSM and a MFCC encoder suggesting that adaptation schemes can be used in DSR applications to increase the robustness of the speech recognition. The additional advantage of the proposed scheme is that it will involve almost no increase in complexity because the adaptation schemes generally have to be used to compensate for other mismatches. Compression distortion (unlike other distortions) is not totally random as we now have knowledge of the distortion introduced (based on the compression scheme) and this can be exploited in the adaptation procedure. Our next goal is to use this information to further improve the recognition performance when compressed data is used for recognition.

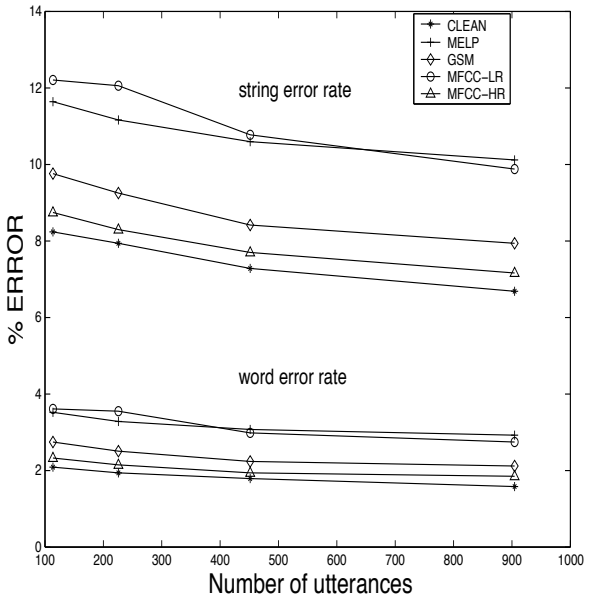

Figure 3: Effect of adaptation data on string error rate and word error rate for clean and encoded data.

6. References

[1] V. V. Digalakis and L. G. Neumeyer, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE Journal on Selected Areas in Communication*, vol. 17, pp. 82–90, January 1999.

[2] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," in *IEEE ICASSP 1998*, pp. 977–980, 1998.

[3] N. Srinivasamurthy, A. Ortega, Q. Zhu, and A. Alwan, "Towards efficient and scalable speech compression schemes for robust speech recognition applications," in *ICME 2000*, July 2000. IEEE International Conference on Multimedia and Expo 2000.

[4] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Tech. Rep. Standard ES 201 108, European Telecommunications Standards Institute (ETSI), April 11 2000.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book (for htk version 3.0)." (htk.eng.cam.ac.uk/prot-docs/HTKBook/htkbook.html), July 2000.

[6] Q. Huo, C. Chan, and C. H. Lee, "Bayesian adaptive learning of the parameters of hidden markov model for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 334–345, September 1995.

[7] N. Srinivasamurthy, A. Ortega, and S. Narayanan, "Efficient scalable speech compression for scalable speech recognition," To appear in *Eurospeech 2001 - Scandinavia*.