A COMPARATIVE STUDY OF HYBRID MODELLING TECHNIQUES FOR IMPROVED TELEPHONE SPEECH RECOGNITION

Rathinavelu Chengalvarayan

Speech Processing Group, Lucent Speech Solutions
Lucent Technologies, Naperville, IL 60566, USA
Email: rathi@lucent.com

ABSTRACT

This paper presents a new technique for modelling heterogeneous data sources such as speech signals received via distinctly different channels which arises when an automatic speech recognition is deployed in wireless telephony in which highly heterogenous channels coexist and interoperate. The key problem is that a simple model may become inadequate to describe accurately the diversity of the signal, resulting in an unsatisfactory recognition performance. To cope up with this problem, different hybrid modelling techniques have been proposed and investigated in this paper by intelligently combining models from two different wireline and wireless environments.

1. INTRODUCTION

A speech signal transmitted through a telephone channel often encounters variable conditions which significantly deteriorate the performance of state-of-the-art HMM-based speech recognition systems [8, 9]. Channel interference and ambient noise are usually the chief contributors to the signal distortion [6, 14]. If no a priori knowledge is provided concerning the nature of the distortion that exists in the network, then acoustic mismatch between the training and the testing conditions would cause a performance degradation that is proportional to the degree of the mismatch [11, 12, 13]. This paper presents a new technique for modelling heterogeneous data sources such as speech signals received via distinctly different channels which arises when an automatic speech recognition is deployed in wireless telephony in which highly heterogenous channels coexist and interoperate.

When speech recognizers are deployed in telephone services, they often encounter variable transmission and background noise conditions, which significantly deteriorate their performance level [14]. To account for the variability due to transmission and noise, we consider multi level cepstral mean subtraction (CMS) techniques [11, 15]. CMS is a standard channel compensation techniques which can remove the time-invariant parts of channel distortion [3]. The effectiveness of CMS is severely limited when the environment can't be adequately modelled by a linear channel [12]. In order to process the non-linear channel, the two level CMS method (2L-CMS) is proposed, where separate channel compensation is performed for segments that are classified as speech and for segments classified as background, and further the system performance depends on the signal classification accuracy [4]. In this paper, we consider the 2L-CMS technique to compensate for the changes in means of the parameters at the feature level. These solutions allow a noticeable recognition error reduction [8].

2. HYBRID MODEL ARCHITECTURES

Different homogenous and heterogenous models were built with same number of Gaussian mixtures as follows. Note that the total number of Gaussian mixtures per model structure is approximately 7072, so that the system complexity remains the same irrespective of model architectures.

- *Wireline*: A separate wireline models were created using wireline data alone.
- *Wireless*: A separate wireless models were trained using wireless data alone.
- *Hybrid-I*: Wireline and wireless models were built separately and combined together with the same model complexity as in *Wireline* and *Wireless* models. The decoder picks up either wireline or wireless models throughout the decoding path depending upon the initial silence classsification as shown in Table 1. That is, if the initial silence is classified as wireless silence then the decoder picks up the wireless models alone and if the initial silence is classified as wireline then the wireline models alone are used for decoding purposes. We also call this model as *homogenous model*, since the decoder path depends upon the initial silence

Sequences	Viterbi Segmentation
Digit String	sil →44 →sil →9Z2 →sil →4213 →sil
Model Path	l → ll → l → lll → l → llll → l
Digit String	sil →Z2 →sil → 593 →sil → O341 →sil
Model Path	w →ww →w →www →w →wwww →w

Table 1. Illustration of Viterbi segmentation using hybrid-I network architecture: 'l' indicates the wireline models, 'w' represents the wireless models and 'sil' is the corresponding silence.

Sequences	Viterbi Segmentation
Digit String	sil →9O1 →sil →761 →sil →8718 →sil
Model Path	l →lww → w →llw → l →lwwl → l
Digit String	sil →34 →sil →Z22 →sil →3829 →sil
Model Path	l → ll → l → lll → l → llll → l
Digit String	sil → 81 →sil → 187 →sil → 8743 →sil
Model Path	w →ww →w →www →w →wwww →w

Table 2. Illustration of Viterbi segmentation using hybrid-II network architecture: 'l' indicates the wireline models, 'w' represents the wireless models and 'sil' is the corresponding silence.

or background classification.

- *Hybrid-II*: Same as previous model structure but the decoder picks up the best model (either wireline or wireless) for a given utterence from an unknown channel as illustrated in Table 2. We call this model *heterogenous model*, since each model has two different pronunciation or variability.
- *Hybrid-III*: A hybrid model was built by using both wireless and wireline training data.

3. FEATURE EXTRACTION

The speech input is sampled at 8kHz and preemphasized using a first-order filter with a coefficient of 0.95. The samples are blocked into overlapping frames of 30 msec in duration, where the overlap is set to 20 msec. Each frame is windowed with a Hamming window and then processed using a 10th-order LPC analyzer. The LPC coefficients are then converted to cepstral coefficients, where only the first 12 coefficients are retained. The basic recognizer feature set consists of 36 features that includes the 12 liftered cepstral coefficients and their first and second order derivatives [2]. Besides the cepstral based features, the normalized energy and its first and second order time derivatives are also computed. Thus, each speech frame becomes represented by a vector of 39 features. Note that the computation of all

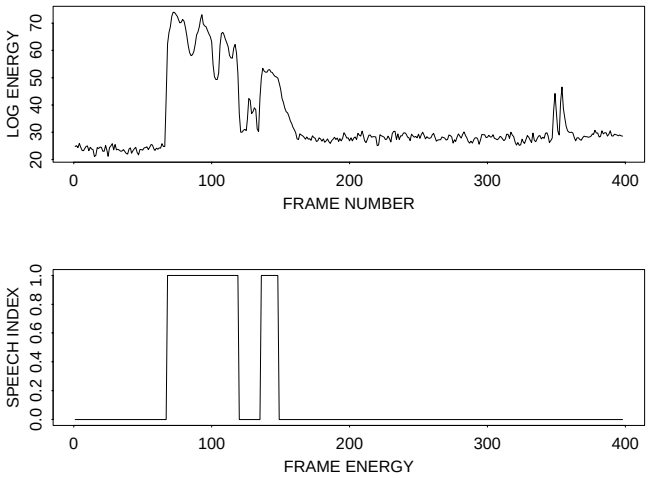

Figure 1. Typical energy measurement contours for the utterance "O182". The top plot show s the original speech energy and the bottom plot shows the speech classification.

the higher order coefficients is performed over a segment of five frames. Since the signal has been recorded under various telephone conditions and with different transducer equipment, each cepstral vector was further processed using the two-level cepstral mean subtraction (2L-CMS) method in order to reduce the effect of channel distortion [4]. The 2L-CMS technique is implemented in several steps:

- Determine the maximum frame energy E_{max} and minimum frame energy E_{min} for every utterence.
- Separating the frames of current utterance into two classes: if $E_t < \alpha \times E_{max} + (1-\alpha)E_{min}$, then the frame t belongs to class-I (silence class), else to class-II (speech class), where α is a constant determined by experiment.
- The background and the speech cepstral mean vectors are calculated for the whole uttereance.
- Finally the normalized cepstral features for each frame are computed by subtracting them by their respective cepstral means.

The above procedure is applied in both training and recognition [2]. To illustrate the nature of the signal classification, Figure 1 shows the actual frame energy trajectory and the corresponding speech index for the connected digit "O182" spoken by a male speaker. It is observed that the 2L-CMS provides better speech and silence classification and further enhances the system performance.

4. SPEECH DATABASE

This section describes the database, LSS_CD, used in this study. This database is a good challenge for speech recognizers because of its diversity. It is a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments. These independent databases are denoted as DB1 through DB6. The LSS_CD database contains the English digits *one* through *nine*, *zero* and *oh*. It ranges in scope from one where talkers read prepared lists of digit strings to one where the customers actually use an recognition system to access information about their credit card accounts. The data were collected over wireline network channels using a variety of telephone handsets. Digit string lengths range from 1 to 16 digits. The LSS_CD database is divided into two sets: training and testing. The training set, DB1 through DB3, includes both *read* and *spontaneous* digit input from a variety of network channels, microphones and dialect regions. The testing set is designed to have data strings from both matched and mismatched environmental conditions and includes all six databases. All recordings in the training and testing set are valid digit strings, totaling 7282 and 13114 strings for training and testing, respectively. Wireless database contains connected digit strings recorded over analog AMPS and digital cellular channels. The collected wireless data include different channel and noise conditions varying from clean speech to hardly audible speech, contaminated mainly by environmental car noise. The digit string length in the wireless database ranges from one to thirty digits. The LSS_CD wireless database used in the experiments is divided into 15488 strings for training and 9142 strings for testing.

5. HMM RECOGNIZER

Following feature analysis, each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models [10]. Each word in the vocabulary is divided into a head, a body, and a tail segment. To model inter-word coarticulation, each word consists of one body with multiple heads and multiple tails depending on the preceding and following contexts. In this paper, we model all possible inter-word coarticulation, resulting in a total of 276 context-dependent sub-word models. Both the head and tail models are represented with 3 states, while the body models are represented with 4 states, each having multiples of 4 mixture components. Silence is modeled with a single state model having 32 mixture components. This configuration results in a total of 276 models, 837 states and approximately 7072 mixture components for wireline, wireless and hybrid models. Training included updating all the parameters of the model, namely, means, variances and mixture gains using ML estimation followed by six epochs of MSE to further refine the estimate of the parameters [5, 7]. Note that the hybrid-I and hybrid-II models have the same set of models since the difference is only in the decoding process. But the parameter complexity of wireline, wireless models used in hybrid-I and hybrid-II is half of that used in actual wireline and wireless models. That is the wireline, wireless models used in hybrid-I and hybrid-II were built with 3536 mixtures respectively, to account for a grand total of 7072 mixture components. So it is now easy to compare the performance with same number of model complexity. The number of competing string models was set to four and the step length was set to one during the model training phase. Each training utterance is signal conditioned by applying 2L-CMS prior to being used in MSE training. The length of the input digit strings are assumed to be unknown during both training and testing.

Type of Model	Wireline Data	
	Word Error	String Accuracy
Wireline	1.138%	94.14%
Wireless	2.593%	88.67%
Hybrid-I	1.442%	92.78%
Hybrid-II	1.148%	94.11%
Hybrid-III	1.108%	94.49%

Table 3. Word error rate and string accuracy for an unknown-length grammar-based wireline connected digit recognition task using the MCE trained wireline, wireless and Hybrid models.

6. RECOGNITION EXPERIMENTS

We have conducted experiments to verify the effectiveness of the proposed hybrid techniques using the continuous speech database on both wireline and wireless connected digit recognition performance. The Table 3 and Table 4 present the word error rates and string accuracy for all the hybrid models. We see that the wireline models behave better than the wireless models for wireline data and the wireless models perform better than the wireline models for wireless data. We can clearly see the mismatch between the two different environments. The hybrid-III model performs better than all other models for wireline data and perform more or less same as in wireless models for wireless data. Similarly hybrid-II model behaves same as the Hybrid-III models, but slightly worse than that of the matched model

Type of Model	Wireless Data	
	Word Error	String Accuracy
Wireline	3.661%	88.04%
Wireless	1.834%	93.94%
Hybrid-I	1.993%	93.67%
Hybrid-II	1.845%	93.86%
Hybrid-III	1.805%	93.99%

Table 4. Word error rate and string accuracy for an unknown-length grammar-based wireless connected digit recognition task using the MCE trained wireline, wireless and Hybrid models.

performance on matched data. But the Hybrid-I model behaves the worst since the initial classification of the silence may not be the correct way of classifying the environment. It is observed that the hybrid-III model outperforms other techniques and exhibits consistent improvements on both wireline and wireless databases. The major benefit of using the hybrid models is that there is no need to know about the source of the data or prior knowledge about the environment.

7. CONCLUSIONS

Different hybrid modelling techniques have been proposed and investigated in this paper by intelligently combining models from two different wireline and wireless environments. The main conclusion is that a single hybrid model is more than sufficient to cater for both wireless and wireline environments without performance degradation. It is generally observed that this kind of hybrid techniques can offer an opportunity for more flexible modelling of speech signals and more sophisticated training of model parameters for speech recognition over diverse telephone networks.

Acknowledgements

The author would like to thank Anand Setlur for his ideas and support in the early stages of this work.

REFERENCES

[1] W. Chou, C.H. Lee, B.H. Juang, and F.K. Soong, "A minimum error rate pattern recognition approach to speech recognition," *International Journal on Pattern Recognition and Artificial Intelligence,* Vol. 8, No. 1, 1994, pp. 5-31.

[2] R. Chengalvarayan, "On the use of normalized LPC error towards better large vocabulary speech recognition systems", *Proc. ICASSP*, 1998, pp. 17-20.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 29, No. 2, 1981, pp. 254-272.

[4] J. Han, M. Han, G.B. Park, J. Park and W. Gao, "Relative Mel-frequency cepstral coefficients compensation for robust telephone speech recognition", *Proc. EUROSPEECH*, 1997, pp. 1531-1534.

[5] B.H. Juang and L.R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 38, No. 9, 1990, pp. 1639-1641.

[6] B.H. Juang, "Speech recognition in adverse environments", *Computer Speech and Language*, No. 5, 1991, pp. 275-294.

[7] B.H. Juang and S. Katagiri, "Discriminative learning for minimum error rate training," *IEEE Transactions on Signal Processing*, Vol. 40, 1992, pp. 3043-3054.

[8] L. Karray, A.B. Jelloun and C. Mokbel, "Solutions for robust recognition over the GSM cellular network", *Proc. ICASSP*, 1998, pp. 261-264.

[9] F. Korkmazskiy, B.H. Junag and F. Soong, "Generalized mixtures of HMMs for continuous speech recognition", *Proc. ICASSP*, 1997, pp. 1443-1446.

[10] C.H. Lee, W. Chou, B.H. Juang, L.R. Rabiner and J.G. Wilpon, "Context-dependent acoustic modelling for connected digit recognition," *Proc. ASA*, 1993.

[11] C.H. Lee, "On stochastic feature and model compensation approaches for robust speech recognition", *Proc. ESCA Workshop on Robust Speech REcognition*, Pont-a-Mousson, France, 1997.

[12] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monne, J. Simonin and K. Bartkova, "Towards improving ASR robustness for PSN and GSM telephone applications", *Speech Communication*, Vol. 23, 1997, pp. 141-159.

[13] J.B. Puel and R. Andre-O'brecht, "Cellular phone speech recognition: Noise compensation versus robust architectures", *Proc. EUROSPEECH*, 1997, pp. 1151-1154.

[14] M. Rahim, B.H. Juang, W. Chou and E. Buhrke, "Signal conditioning techniques for robust speech recognition", *IEEE Signal Processing Letters*, Vol. 3, No. 4, 1996, pp. 107-109.

[15] O. Viikki, D. Bye, K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise", *Proc. ICASSP*, 1998, pp. 733-736.