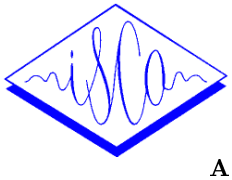
USE OF MULTIPLE CLASSIFIERS FOR SPEECH RECOGNITION IN WIRELESS CDMA NETWORK ENVIRONMENTS



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Rathinavelu Chengalvarayan

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

ABSTRACT

In this paper, we address the problem and the use of multiple classifiers for robust recognition over the cellular network. The idea is to provide more variability to the system to be trained, and to support this variability with more number of model parameters. The main drawback is that the model size, and the computational complexity increases linearly related to different call environment. To alleviate this problem we first introduce a new measure called the average-arc-count into the decoding process. The main advantage of this new measure is that many of the multiple classifiers can be shut down during the recognition stage if the average-arc-count of individual classifier exceeds a certain threshold limit for a given utterance. Secondly, we can also build individual classifiers with less number of parameters and without degrading the overall system performance. Experimental results on English connected digit recognition task show a string error rate reduction of as much as 40% by using the multiple classifiers when compared to individual CDMA systems.

1. INTRODUCTION

When speech recognizers are deployed in cellular networks, they often encounter variable transmission and background noise conditions, which significantly deteriorate their performance level [6]. Although significant improvement is obtained by a suitable choice of the spectral representation, model transformation and the associated distance measure, the problem is not completely solved [5]. Another intuitive direction to increase the immunity of the recognizer to ambient noise consists of filtering out the noise from the recorded signals [4]. A popular technique in speech enhancement applications is spectral subtraction. Moderate results are obtained using linear and nonlinear spectral subtraction, which produces musical noise at the filter output [9]. This resulting musical noise needs to be cancelled prior to performing ASR. Not doing noise cancellation prior to ASR would result in the musical noise falsely triggering the recognition system [7].

In this paper, we address the problem and the use of multiple classifiers for robust recognition over the cellular network. The idea is to provide more variability to the system to be trained, and to support this variability with the greatest number of parameters. We have conducted experiments to verify the effectiveness of the proposed multiple classifier using the noisy speech database on both landline and cellular CDMA connected digit recognition performance. Incorporation of muliple classifiers is observed to lead to about 40% reduction in errors. Further evaluation of the multiple modeling technique was conducted using whole word model for CDMA data in order to reduce the model complexity without degrading the overall system. It is encouraging that our goal of designing a single global system for all three networks (landline, wireless and cellular CDMA) can be achieved by using the multiple classifiers, and the test results demonstrate the efficacy of built-in shut down early capability using the new measure called *average-arc-count* into the decoding process. The additional benefit from a shut down early procedure is that the recognition resource is freed up earlier to process the next request. This means that we can, on average, process more calls with less computational resources.

2. HMM-BASED HYBRID SYSTEMS

To take more variable conditions into account, it is convenient to increase the training set to include all the network specific databases (analog, digital, wireless and cellular). Different homogenous and heterogenous models were built with same number of Gaussian mixtures as follows.

- *Baseline Hybrid System*: Earlier we proposed a hybrid modeling scheme, which was shown to reduce error rates in noise robust telephone speech recognition [2]. This new baseline system is built by combining both the landline and wireless databases together during HMM training process and with half of the parameters as in our earlier system [2].
- *Basic Hybrid System*: Hybrid model is trained with data recorded through landline, wireless and CDMA cellular networks.
- *CDMA System*: A separate CDMA model is created by using only the CDMA (handset, lapel and visor) data during training.
- *Enhanced Hybrid System*: HMM training is carried out on increasing set of speech recordings. This system is trained with a global data by intelligently combining data from many different landline, wireless and cellular networks. We also added some artificially simulated CDMA data to incorporate and manage more variability in the training data. In all our experiments, we performed 1-bit clipping on all CDMA training data and used them during the regular model building process.

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 multiple silence network architecture: 'l' indicates the baseline hybrid models, 'w' represents the CDMA models and 'sil' is the corresponding silence.

The main motivation is that the simulated 1-bit clipping effect on the training set looks similar to the natural clipping on testing data. Thus, the models trained with such simulated data might as well improve the recognition performance on the unknown testing data clipping effect. We can also include the two or more bit clipped CDMA data during training process, but we didn't explore this option in the current study.

Note that the total number of Gaussian mixtures per model structure is approximately 3904, so that the system complexity remains the same irrespective of model architectures.

3. HMM-BASED MULTIPLE CLASSIFIERS

Different homogenous and heterogenous models were built with 7808 number of Gaussian mixtures as follows.

- *Multiple Silence System*: Baseline hybrid and CDMA models were combined together with double the model complexity as in any *hybrid* models. The decoder picks up either baseline hybrid or CDMA 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 CDMA silence then the decoder picks up the CDMA models alone and if the initial silence is classified as baseline hybrid then the baseline hybrid models alone are used for decoding purposes. We also call this model as *homogenous model*, since the decoder path depends upon the initial silence or background classification [2].
- *Multiple Pronunciation System*: Same as previous model structure but the decoder picks up the best model (either baseline hybrid or CDMA) 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 [2].
- *Multiple Classifier System*: The baseline hybrid and CDMA recognizers will run in parallel during decoding, and the system picks the best result based on average log-likelihood scores of individual systems as shown in Figure 1.
- *Reduced Multiple Classifier System*: Further evaluation of the multiple classifier system using whole word model for CDMA data was conducted. This system is same as multiple classifier system but with whole-word models for CDMA and head-body-tail models for baseline hybrid. The model size is about 25% more than the individual hybrid systems.

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 multiple pronunciation network architecture: 'l' indicates the baseline hybrid models, 'w' represents the CDMA models and 'sil' is the corresponding silence.

- *Enhanced Multiple Classifier System*: When the landline data is tested on a CDMA system, the average-arc-count increases tremendously to a point where the recognizer is overloaded with unnecessary log-likelihood computations, arc-expansions, etc. which results in a longer delay in reporting the recognized string. This is true to some extent that the CDMA models look more fuzzier when tested on mismatched landline data and hence the average-arc-count gets incremented and slow down the decoding speed [1]. To overcome this problem, methods for extending the multiple classifier system to real-time implementation is illustrated in Figure 2. When the input speech belongs to landline environment then the baseline hybrid recognizer will become more active and report the N-best results to candidate picking module. The CDMA systems will be shut down early due to the higher average-arc-count measure and the recognizer resources will be released to other waiting speech input. Then the best candidate will be selected based on the baseline hybrid recognizer output. However, if both the recognizers are active during the decoding process then the candidate with higher log-lokelihood and with lower average-arc-count will be selected as the recognized string. If the average-arc-count exceeds certain threshold limit on both the classifier then that speech input is considered as rejection.

4. SPEECH DATABASES

The data distribution of the training and testing set is shown in Table 3. The English database contains digits one through nine, zero and oh. Digit string lengths range from 1 to 30 digits. The digit string length is fixed to 10 for CDMA training and testing data as well as for landline testing data. The landline and wireless data are a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments [2].

The CDMA corpus comprised speech data from 127 speakers (41 male and 59 female speakers for training and 16 male and 11 female speakers for testing) collected in a car driving on different highways at 55 mph or greater. The windows were closed and the radio and fan were partially switched off. The subject was seated in the front passenger side and the data were recorded on CDMA cellular network via typical hands-free microphones: handset (handheld), lapel and visor-mounted (half-duplex). The motivation for

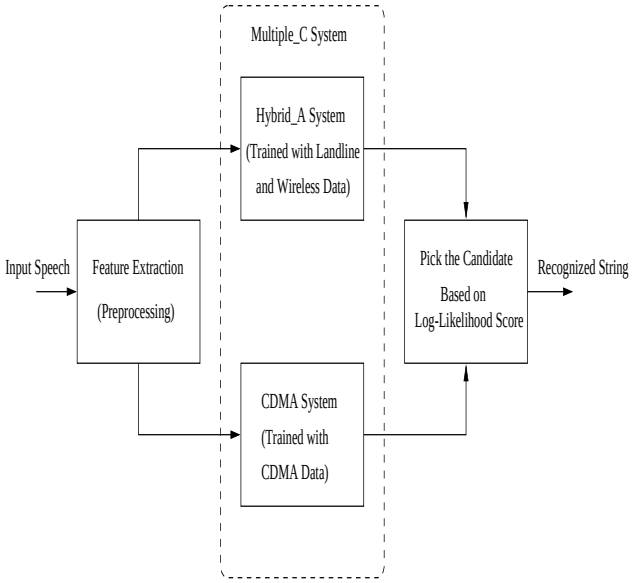


Figure 1. A block diagram of a typical parallel speech recognition system using multiple classifiers.

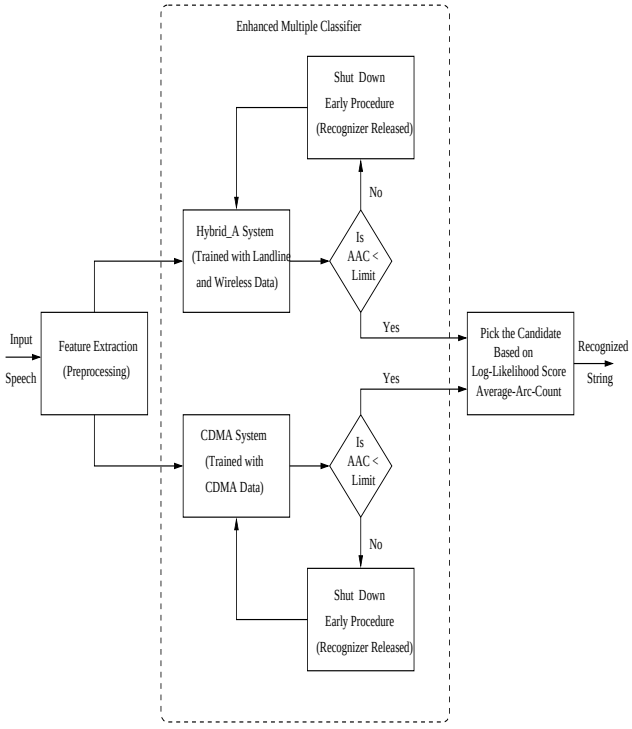


Figure 2. A block diagram of an enhanced multiple classifier system incorporating the shut down early procedure.

Databases	Training		Testing	
	Str	Spk	Str	Spk
Landline (Digital)	24378	$\approx$ 5000	2203	$\approx$ 500
Wireless (AMPS & TDMA)	15070	$\approx$ 2500	—	—
CDMA (Handset)	769	100	256	27
CDMA (Lapel)	2026	100	517	27
CDMA (Visor)	1880	100	477	27
CDMA (Simulated)	4675	100	—	—

Table 3. Distributions of spoken digit strings and the speaker population among the training and testing sets of the connected digit database.

adding more simulated data in the training set is that about 20% of the handsfree data is partially clipped and also the collection of large CDMA database is very expensive and time consuming [11]. The simulation of above clipping can be done on CDMA training data by first converting the 8-bit μ-law samples into a 16-bit linear digital format. Then the samples are multiplied by 2 and clipped at their minimum or maximum allowed value (-32768 or 32767). We call this type of clipping as 1-bit clipping and similarly the n-bit clipping can be done recursively [3].

5. EXPERIMENTAL RESULTS

The recognizer feature set consists of 39 features that includes the 12 liftered cepstral coefficients, log-energies, their first and second order derivatives [2]. The energy feature is batch normalized during training and testing [3]. 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. In this study, we model all possible inter-word coarticulation, resulting in a total of 276 context-dependent sub-word models. Each model is represented with 3 or 4 states, each having multiples of 4 mixture components. Silence is modeled with a single state model having 32 mixture components [3]. For the whole word CDMA system, each digit was modeled with 10 state HMMs, with 8 Gaussian mixtures. Training included updating all the parameters of the model, namely, means, variances and mixture gains using a few epochs of minimum string error training [3]. Each training utterance is signal conditioned by applying batch-mode cepstral mean subtraction prior to being used in training [10].

The Table 4 presents the string accuracy for baseline (Hybrid_A), basic (Hybrid_B), and enhanced (Hybrid_C) hybrid systems along with a separately trained CDMA alone system. We observed that the Hybrid_A model behaves better than the CDMA model for landline data and the CDMA model performs better than the Hybrid_A model for CDMA data. We can clearly see the mismatch in performance between the two different environments. Hybrid_B is better than the CDMA model on landline data and provides a substantial improvement on CDMA data when compared to

Type of Database	Type of Model			
	CDMA	Hybrid_A	Hybrid_B	Hybrid_C
Handset	90.2%	93.8%	95.7%	94.6%
Lapel	86.5%	76.7%	84.0%	87.4%
Visor	65.4%	36.1%	47.6%	54.3%
Landline	79.9%	92.7%	91.9%	92.7%
Overall	79.6%	82.6%	84.9%	86.7%

Table 4. String accuracy for a 10-digit grammar-based English connected digit recognition task as a function of hybrid model type.

Type of Database	Type of Model			
	Multi_A	Multi_B	Multi_C	Multi_D
Handset	88.3%	90.2%	86.3%	89.8%
Lapel	80.7%	87.3%	83.8%	86.5%
Visor	85.0%	65.4%	62.9%	65.4%
Landline	87.1%	87.0%	92.6%	92.7%
Overall	81.4%	84.3%	86.7%	87.8%

Table 5. String accuracy for a 10-digit grammar-based English connected digit recognition task as a function of multiple classifier type.

Hybrid_A system. Overall Hybrid_C model performs better than the CDMA model on handset and lapel data and not on visor data. The performance of Hybrid_C model remains the same as Hybrid_A on landline data. On average, we noticed a 13% string error rate reduction by using the Hybrid_B when compared to Hybrid_A (improvement from 82.6% to 84.9%). Further reduction of 11% string error rate is obtained by using the more enhanced Hybrid_C model (improvement from 84.9% to 86.7%).

Further test results using multiple classifier (Multiple_D) are tabulated in Table 5 together with the multiple classifier using whole word model for CDMA data (Multiple_C), a system using multiple pronunciation models (Multiple_B) and multiple silence models (Multiple_A). Multiple_A system yields the worst since the initial classification of the silence may not be the correct way of classifying the environment. Multiple_B is better than the Hybrid_A and Hybrid_B models but inferior to more enhanced Hybrid_C system. We observed that the multiple pronunciation for individual words in the lexicon may not be the right choice in accelerating the system robustness. Overall Multiple_D outperforms all the other models and yields about 40% in string error rate reduction when compared to the CDMA system. Multiple_D exhibits consistent improvements on both landline and CDMA databases. When moving from head-body-tail to a whole word model structure for the CDMA system, the Multiple_C dropped almost an absolute percentage point when compared to Multiple_D system. On the other hand, the model size of Multiple_C is about 40% less than the Multiple_D system. Preliminary recognition results showed similar string accuracy as that of Multiple_D and resulted in a 35% reduction in real-time using the average-arc-count based shut down early procedure. It is encouraging that our goal of designing a single global system for all three networks (landline, wireless and cellular CDMA) is achieved by using the Multiple_D, and the test results have demonstrated the efficacy of enhanced multiple hybrid systems.

6. CONCLUSIONS

In this paper, we described the problem and the use of multiple classifiers for robust recognition over the cellular network. The average-arc-count measure was introduced into the decoding process to minimize the computation complexity. Further evaluation of the multiple classifier system using whole word model for CDMA data was explored with 40% reduction in model size when compared to the top-performing multiple classifier system. The multiple classifier performance is significantly better than the previously proposed hybrid system built with data taken from all three networks [2, 3]. Despite the recent progress, robustness is still a major research issue. The proposed multiple classifier design algorithm is extremely useful in various aspects of speech recognition, ranging from language identification [3] to robust design of dialogue systems [8].

REFERENCES

[1] E. Burhke, W. Chou and Q. Zhou, "A wave decoder for continuous speech recognition", *Proc. ICSLP*. pp. 2135-2138, 1996.

[2] R. Chengalvarayan, "A comparative study of hybrid modelling techniques for improved telephone speech recognition", *Proc. ICSLP*, pp. 313-316, 1998.

[3] R. Chengalvarayan, "Hybrid HMM architectures for robust speech recognition and language identification", *Proc. Systemics, Cybernetics and Informatics*, Vol. 6, pp. 5-8, 2000.

[4] Eric J. Diethorn, "A subband noise-reduction method for enhancing speech in telephony and teleconferencing", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[5] Y. Gong and J. J. Godfrey, "Transforming HMMs for speaker-independent hands-free speech recognition in the car", *Proc. ICASSP*, pp. 297-300, 1999.

[6] J-C. Junqua, "Impact of the unknown communication channel on automatic speech recognition: A review", *Proc. EUROSPEECH*, pp. 29-32, 1997.

[7] C. Mokbel and G. Chollet, "Automatic word recognition in cars", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, pp. 346-356, 1995.

[8] P. Niyogi, J-B. Pierrot and O. Siohan, "Multiple classifiers by constrained minimization", *Proc. ICASSP*, 2000.

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

[10] 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, pp. 107-109, 1996.

[11] C. Tarcisio, F. Daniele, G. Roberto and O. Marco, "Use of simulated data for robust telephone speech recognition", *Proc. EUROSPEECH*, pp. 2825-2828, 1999.