ROBUST SPEECH RECOGNITION FOR MULTIPLE TOPOLOGICAL SCENARIOS OF THE GSM MOBILE PHONE SYSTEM

Theodoros Salonidis, Vassilios Digalakis
Technical University of Crete
Kounoupidiana, Chania GREECE
email : {thsalon,vas}@telecom.tuc.gr

ABSTRACT

This paper deals with robust speech recognition in the GSM mobile environment. Our focus is on the voice degradation due to the losses in the GSM coding scheme. Thus, we initially propose an experimental framework of network topologies that consists of various coding-decoding systems placed in tandem. After measuring the recognition performance for each of these network scenarios, we try to increase recognition accuracy by using feature compensation and model adaptation algorithms. We first compare the different methods for all the network topologies assuming the topology is known. We then investigate the more realistic case, in which we don't know the network topology the voice has passed through. The results show that robustness can be achieved even in this case.

1. INTRODUCTION

Long distance telephone connections are established using a multitude of terrestrial or wireless telecommunication circuits. Each one of them utilizes a different coding scheme and a different coding rate and a telephone session is formed by the tandeming of these different systems. In the past there have been several efforts to evaluate the voice quality of such tandem connections and scenarios using the subjective 5-point Mean Opinion Score (MOS) transmission quality measure [1][2]. However, we focus on the problem of having a speech recognizer at the receiver instead of a human being, where the different tandeming connections affect recognition accuracy dramatically.

We have investigated recognition performance in a number of different network configurations of the GSM environment. Each network scenario corresponds to a different acoustic environment that the recognizer must deal with. These scenarios were simulated in software and comprised transport over ITU-T G.711 64 Kbps PCM channels, G721 32kbps and G.723 16 kbps ADPCM and full rate RPE-LTP GSM 13 kbps speech encoders.

Scenario 1 refers to three GSM RPE-LTP encoders connected in tandem. We assume that this is the maximum number of GSM encoders that can be placed in tandem and this schema yields the greatest degradation (worst case). Indeed in practice, it has been shown that careful network planning should avoid placing more than 3 GSM encoders in tandem [8]. Topology 2 reflects the case where both transmitter (user) and receiver (recognizer) utilize GSM encoding. Topology 3 refers to the simple case where the user is mobile and the recognizer is accessed via conventional wireline connection (PCM). In international telephone connections Digital Circuit Multiplication Equipment (DCME) technology based on ITU-T G.726 32 Kbps ADPCM codecs is used [3]. Topology 4 is the case of an international call where we have DCME 16kbps for the international network, DCME 32 Kbps for the national network and finally GSM termination for the receiver. Topologies 5 and 6 are simpler cases where we employ only 32 kbps DCME technology for the international network.

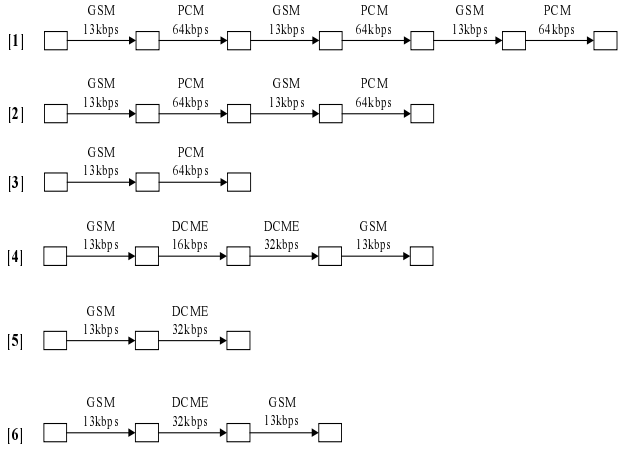

Figure 1: Network topologies for the tandeming of various encoders

2. COMPENSATION TECHNIQUES

When the recognizer is trained under a different acoustic environment than the one in which it is being tested, the mismatch affects its recognition accuracy. To overcome this acoustic mismatch, several compensation techniques have been devised. In our investigation we used two approaches, namely the Probabilistic Optimum Filtering (POF) algorithm [4] and adaptation techniques [5]-[7].

The POF algorithm is a mapping algorithm that is trained by using simultaneous recordings of noisy and clean data. After training, the mapping is applied to incoming noisy speech feature vectors to produce estimated "clean" speech feature vectors that are recognized by the recognizer afterwards.

Adaptation techniques are "model oriented", that is they attempt to adapt the recognizer model parameters to the incoming noisy

data. We assume that the vector process $\{\mathbf{y}_t\}$, that corresponds to the new acoustic GSM environment, can be obtained from $\{\mathbf{x}_t\}$, the stochastic process that corresponds to the initial "clean" model, by the relation:

$$\mathbf{y}_t = \mathbf{A}_g \mathbf{x}_t + \mathbf{b}_g ,$$

where the $\mathbf{A}_g, \mathbf{b}_g, g = 1,\ldots, N_g$ are matrices used to transform the means and variances of the Gaussians of the initial ("clean") model. (N_g is the number of transformations). The index *g* depends on the hidden Markov model (HMM) state used at time *t*. We used this technique in two different variations:

Method I: In this method [5], the matrix $\mathbf{A}_g$ is diagonal, and is applied to both the means and variances of the HMM Gaussians.

Method II: Here, the matrix $\mathbf{A}_g$ is block diagonal, and only the means are transformed while the variances remain the same (MLLR, [7]).

3. EXPERIMENTS

The recognizer used throughout our experiments is SRI's DECIPHER phonetically tied-mixture speech recognition system [9-10]. The signal processing consists of a filterbank-based front-end that generated six feature streams: the cepstrum (c1-c8), the cepstral energy (c0), and their first and second order derivatives. The dimension of the final feature vector is 27 (3*9) because we are dealing with telephone bandwidth data. To evaluate our algorithms we used the air-travel information domain (ATIS) with data collected over the telephone network. A bigram language model was used for all our experiments.

The POF algorithm was trained by simultaneous recordings of PCM 64-kbit speech (this is referred as the "clean speech") and "noisy speech", that has passed from the various tandeming network topologies. The same set of training data was used in the model adaptation techniques.

The initial models of the recognizer have been trained with clean speech data using a large collection of hundreds of thousands sentences taken from several databases. The POF mapping models were trained using a set of 400 sentences of male-female speakers, taken from SRI's stereo ATIS database. The test set consists of 210 sentences taken from the same database.

As a performance indicator for our recognizer we used the *Word Error Rate*, that is the percentage of words that were 'erroneously' recognized. 'Erroneously' means that the recognizer has added, deleted or replaced some of the words that have been spoken in the initial sentence. Thus:

$$WER = \frac{INS + DEL + SUB}{TOTAL} \times 100\%$$

3.1. POF Experiments

We first measured the recognition error in the matched condition, that is the error of the recognizer trained on clean speech and tested on clean speech. We also measured the recognition error for the mismatched case of different scenarios, which is the error of the "clean" recognizer on speech data which have passed from this scenario. These error rates can be considered as lower and upper bounds in the recognition performance of any compensation algorithm, and for the GSM3 (topology 1) data are summarized in Table 1.

	Test Data	
	Clean	GSM3
(%Error)	13.30	23.75

Table 1: Lower and upper recognition bounds for the GSM3 topology

Given these limits, we tried to apply the compensation techniques in order to improve the recognition performance in the mismatched case of the GSM3 topology.

The parameters controlling the POF algorithm were actually the *number of Gaussian* distributions that define the number of the VQ regions used in the mapping algorithm (actually each Gaussian models one of the VQ regions that comprise the acoustic space), and the *delay* that defines the number of neighbouring frames of the noise frame $\mathbf{Y}_n$ that are taken into account when estimating the clean vector $\hat{\mathbf{x}}_n$.

POF compensation results			
	# of Gaussians		
Delay	**5**	**10**	**50**
0	19.3	18.9	18.8
1	18.0	18.2	18.2
2	18.8	18.7	19.0

Table2 : Recognition error rates for the GSM3 topology using the POF algorithm

We see that the error does not decrease as the number of Gaussians (VQ regions modeling the acoustic space) increases. Moreover, the increasing delay does not seem to affect the recognition error. Similar results were obtained for the other topologies that we examined. Our conclusion is that the GSM noise does not follow some specific noise pattern that the POF algorithm can model efficiently by using more Gaussians. Even if the recognition error generally decreased from the 23.75% upper limit, it was not possible to find a specific pair of parameters (delay, #of Gaussians) that yield the best result in every case.

3.2. Adaptation Experiments

After POF, we used the transform adaptation method on the same training and test data with POF. We used variations I and II for

$N_g = 2,5,10,20,30,41$. The results for the GSM3 data, are summarized in Table 3. We observe that method I outperforms method II. The reason is that method I changes the variances of the model's Gaussian mixtures. The transform matrix $\mathbf{A}_g$ for method I had coefficients greater than unity, indicating that the variances of the adapted models were increased.

Transform Adaptation results						
	Number of transforms					
Meth.	2	5	10	20	30	41
I	18.7	18.3	**17.3**	17.8	18.7	19.1
II	20.6	20.6	21.3	22.4	22.0	22.0

Table 3: Recognition error rate for adaptation methods I, II (MLLR) and for different numbers of transformations.

This corresponds to smoother phoneme representations of the adapted HMM models, rendering them more robust to the GSM noise. Moreover, by observing the $\mathbf{A}_g$ matrix, some phoneme classes benefited from a greater increase in the Gaussians' variance. This effectively shows that certain identifiable classes of phonemes are more severely affected from the noise in GSM encoding. Since variance compensation plays such a significant role, MLLR didn't work well in this case.

Another observation stems from the fact that there is a minimum error rate (17.23%) for $N_g = 10$. Further experimentation with the remaining topologies indicated that $N_g = 10$ is the selection that yields the maximum error rate reduction.

In general, adaptation method I achieved lower error rates than POF, as we can see in Figure 2 which extends the previous experiment to the rest of the network topologies. Topology 3 is omitted, since the uncompensated error rate in this case is 14.5%, not significantly higher compared to the lower limit (13.3%).

The first bar in each topology shows the uncompensated error, the second the error using POF compensation, and the third using method I with $N_g = 10$.

3.3. The "Cocktail" transform

All previous experiments assumed knowledge of the noise environment that caused voice degradation. For each case, we trained our transforms in the corresponding acoustic space. In realistic scenarios, however, we don't know the network topology that caused the degradation. A solution to this problem is to train the recognizer's models with data collected from all six network topologies. Thus, we adapted the initial models of the recognizer using 400 training sentences equally distributed over all six different topologies. The results of this "cocktail" transformation training tested under each topology are summarized in Table 4. We used adaptation method I with $N_g = 10$ transformations.

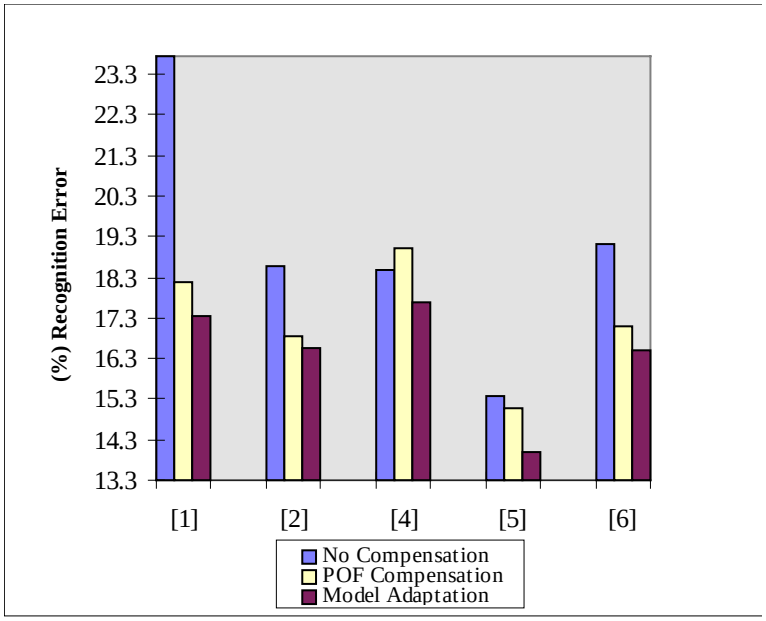

Figure 2 : Cumulative results for all the topologies

The results of the "cocktail" training are surprisingly very close to those that were obtained in the previous "best" case where training and testing environments are matched. This fact provides an even stronger argument for choosing adaptation method I to cancel the GSM noise.

Test / Train	Clean	[1]	[2]	[4]	[5]	[6]
Clean	13.3	23.75	18.55	18.49	15.38	19.14
[1]		17.32				
[2]			16.57			
[4]				17.69		
[5]					14.26	
[6]						16.51
Cocktail	13.19	17.96	15.66	16.94	13.99	17.05

Table 4 : Error rates corresponding to all the topologies with matching training and testing acoustic environments as opposed to the "Cocktail" training results.

Adaptation method I acts in our case as a phone- and feature-dependent variance smoothing scheme. To further elaborate on this, we present in Table 5 the first few diagonal elements of the transformation matrix A for each one of the ten groups of phones that were determined using clustering. Since matrix A pre- and post-multiplies the covariance matrix in adaptation method I, the variances of a particular cepstral coefficient are multiplied by the square of the value of the corresponding element of the transformation matrix. Values of the matrix elemens greater than one imply an increase in variance and smoothing of the corresponding distribution. For example, we can see in Table 5 that the variance of the first cepstral coefficient is increased (and hence, the corresponding distribution smoothed) more for vowels

than for consonants. Similarly, the opposite is true for the fourth cepstral coefficient.

Phone Cluster	C1	C2	C3	C4
s z sh zh ch jh th f hh	0.95	1.03	0.99	1.08
d dh dx t	1.02	1.02	1.03	1.07
b v p	1.02	1.05	1.01	1.05
l el r er	1.04	0.99	0.99	1.02
g k	1.04	1.03	1.01	1.07
w uw aw ow	1.05	0.99	1.00	1.03
aa ao ae ah ax eh ih	1.05	0.99	1.03	1.02
m em n en ng	1.05	1.02	0.98	1.03
y iy oy ey ay	1.08	0.97	0.97	0.97

Table 5: Values of transformation parameters for different cepstral coefficients and different classes of phones.

4. CONCLUSIONS

By applying compensation methods to data passed from a set of network topologies in the GSM environment, we managed to reduce the recognizer error rate. We have shown that the POF method reduces the recognition error rate, but is unstable in terms of a choice of parameters that would yield the optimal results. This leads to the conclusion that the GSM noise does not follow a rich noise pattern and therefore techniques of feature mapping such as POF do not work very efficiently.

Adaptation techniques acting on both the means and the variances of the Gaussians of the recognizer's acoustic models were more efficient than POF in reducing the mismatch using a small number of transformations, $N_g = 10$.

Finally we proposed a "cocktail" transformation method, where the recognizer models were trained using a collection of data from all six topological scenarios. The results tended to be very close to the ones of the case where the training and testing environments are the same. This result shows that the proposed adaptation techniques can be used in real world situations where the network topology is not known at the receiver.

5. REFERENCES

[1] Dimolitsas S. et al: "Voice Quality of Interconnected North American Cellular, European Cellular, and Public Switched Telephone Networks" 45th IEEE Vehicular Technology Conference (VTC '95) Chicago Illinois,USA 1995.

[2] Simao F. Campos Neto, franklin Corcoran & Ara karahisar, "Performance Assessment of tandem Connection of Cellular and Satellite-mobile Coders."

[3] CCITT Recommendation G.763. Digital Circuit Multiplication Equipment Using 32 Kbps ADPCM and Digital Speech Interpolation, Geneva, 1991.

[4] Leonardo Neumeyer and Mitchel Weintraub, "Microphone-Independent Robust Signal Processing using Probabilistic Optimum Filtering", 1994 IEEE ICASSP, pp I417-I420.

[5] V. Digalakis, D. Rtischev & L. Neumeyer, "Fast speaker adaptation using constrained estimation of Gaussian Mixtures". IEEE transacations on Speech and Audio Processing, vol. 4, pp. 294-300, July 1996.

[6] L. Neumeyer, A. Sankar & V. Digalakis, "A Comparative study of speaker adaptation techniques", Proc. Eurospeech, pp. 1127-1130, Madrid, 1995.

[7] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, pp. 171-185, 1995.

[8] B.T. Lilly and K.K. Paliwal, "Effect of Speech Coders on Speech Recognition Performance"

[9] V. Digalakis and H. Murveit, "GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", 1994 IEEE ICASSP, pp I537-I540.

[10] H.Murveit, J. Butzberger, and M. Weintraub, "Large Vocabulary Dictation Using SRI's DECIPHER Speech Recognition System: Progressive Search Techniques", 1993 IEEE ICASSP, pp. II 319-II 322.