A Comparative Study on the Effect of Different Codecs on Speech Recognition Accuracy Using Various Acoustic Modeling Techniques

Srinivasa Raghavan, Nisha Meenakshi, Sanjeev Kumar Mittal, Chiranjeevi Yarra
Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, Karnataka, India
Email: srinivasaraghavankm@gmail.com, gnisha@ee.iisc.ernet.in, rsr.skm@gmail.com, chiranjeevi.yarra@ee.iisc.ernet.in
Anupam Mandal, K.R. Prasanna Kumar
Center for AI and Robotics, Bangalore 560093, Karnataka, India
Email:{amandal,prasanna}@cair.drdo.in
Prasanta Kumar Ghosh
Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, Karnataka, India
Email: prasantg@ee.iisc.ernet.in

Abstract—**In this work, we study the effect of codec induced distortion on the speech recognition performance in the TIMIT corpus using eleven codecs and five acoustic modeling techniques (AMTs) including several state-of-the-art methods. This study is performed in a single round of encoding-decoding and various tandem scenarios. Experiments from the single encoding-decoding case reveal that the acoustic models from G.711A, a narrowband high bit rate codec yields lower phone error rate (PER) compared to low bit rate codecs for most AMTs. It is observed that among the eleven codecs based acoustic models, G.711A, G.728, G.729B, AMR-WB and G.729A codecs consistently result in the least five PERs across AMTs. It is found that the model trained on 'clean' speech data (PCM) performs poorly in three of the five AMTs compared to these five codec based acoustic models. These five models are then used in six different tandem scenarios comprising three unseen codecs. Similar to the single round of encoding-decoding case, the PER for each of the tandem scenarios turns out to be the lowest consistently for all AMTs when the acoustic model from the G.711A codec is used. However, when the acoustic model is trained with mixed speech data from all tandem scenarios, the PER is found to perform better than the matched condition in the case of four out of five AMTs.**

I. Introduction

The ever decreasing cost of mobile and IP telephony has spurred new models of customer service based on remote communication. Such models of service require activities like automatic analysis of user feedback and detection of policy violations to improve customer experience. This is generally done by monitoring the content of voice communication and subsequent analysis by employing automatic speech recognition (ASR) systems. However, speech/voice data on which the recognizers work typically undergoes distortion introduced by the codec and channel. Specifically, while a codec compresses the information in a small number of bits, the task of an ASR system, on the other hand, requires the extraction of discriminative features. To cater to such loss in information, the field of speech recognition has witnessed key improvements on acoustic modeling techniques (AMT) such as discriminative training [1], subspace Gaussian mixture models (SGMM) [2], and deep neural networks (DNN) [3].

The effects of codecs on the performance of speech recognition have been studied earlier [4], [5], [6], [7]. These studies were carried out on decoded speech obtained after a single round of encoding and decoding. Though all these studies reported a drop in the recognition accuracy for codec with reduced bit rate, Lilly et al [5] concluded that the drop was not strictly monotonic. Contrary to the findings by Euler et al [4], the acoustic model trained with pulse coded modulation (PCM) speech was found to yield the best recognition accuracy in the case of GSM codecs [6], [7]. Recognition performance under noisy conditions has also been studied [8] using G.729, G.723.1 and three codecs in the GSM family, namely, GSM-AMR, GSM-HR, and GSM-FR. Another scenario pertaining to the codec induced distortion is tandeming wherein the speech is processed through several coding schemes as it traverses through different media gateways. The adverse effect of tandeming on speech recognition performance has also been studied [9], [5]. Lilly et al [5] concluded that the effect of tandeming on the recognition performance is more for low bit rate codecs (13 kbps or less) compared to the high bit rate codecs (40-64 kbps). A variety of techniques has been proposed in the literature to compensate the drop in recognition performance due to distortions introduced by codecs and tandeming. These techniques include enhancement of the decoded speech, robust feature extraction [10], compensation in feature space [11], augmentation of codec parameters by including extra data containing compensation information [12] and adaptation of acoustic models [13], [14], [9]. However, such compensations can only be partial as the distortions are non-linear and the topologies are generally unknown.

In the light of the new developments in the field of AMTs and codecs, in this work, we study the effect of codec specific distortion on speech recognition performance using the TIMIT database and eleven codecs of different types including narrowband, wideband, parametric, hybrid and waveform coding. To the best of our knowledge, the effects of such codec induced distortions in conjunction with those AMTs on the ASR performance is not yet well understood. Therefore, it is necessary to first study the effects of such codecs when applied to microphone speech data under clean conditions. Hence, we use the TIMIT database in this study so that distortion using different codecs can be simulated and recognition performance can be studied in a controlled manner. Since the focus of the current work is to understand the effects of the codec induced distortions, the effects of channel induced distortions are kept out of the scope of this study. Experiments are performed using five AMTs including GMM-hidden Markov model (GMM-HMM) frameworks as well as several recently developed AMTs such as SGMM and DNN. In order to understand the robustness of different AMTs in the presence of codec induced distortions, we exclude the influence of the language model, in speech recognition, by using a zero-gram language model [15]. Experiments on the speech data from the single round of coding-decoding reveals that acoustic models derived from codecs with higher bit rate exhibit lower phone error rate (PER) irrespective of the AMT. This is consistent with the findings from the literature. The study also reveals that the DNN based AMT not only performs the best in PCM coded speech but also achieves the lowest PER on the decoded speech for each codec.

This study also investigates the effect of tandeming using six unseen tandem scenarios. Similar to the single encoding-decoding case, the PER turns out to be the lowest corresponding to the acoustic model from a codec with a high bit rate. We also consider an acoustic model built using a mixture of speech data from six tandem scenarios, referred to as *cocktail* data. Our study reveals that the acoustic model obtained with *cocktail* data results in a PER lower than that from the matched condition. Thus, in the absence of any knowledge about the tandem topologies, the best strategy could be to use an acoustic model from the high bit-rate codec. However, if the pool of the tandem topologies is known, the acoustic model derived from a cocktail speech data could be used.

II. Selection of acoustic modeling techniques and codecs

A. Selection of acoustic modeling techniques

The AMTs used in the state-of-the-art large vocabulary speech recognition systems (LVCSR) can be grouped into two categories:

1) *GMM-HMM based system* uses GMMs for modeling individual HMM states. Here, the GMMs are used to represent the emission density of an HMM. The training can be done either using maximum-likelihood [16] or sequence discriminative criteria namely Maximum Mutual Information (MMI) [17], boosted MMI [18], Maximum Phone Error (MPE) [19] or Minimum Bayesian Risk (MBR) [20].

2) *DNN-HMM based hybrid system* is the one where the DNN is trained to provide posterior probability estimates corresponding to HMM states. The posterior probabilities are converted into likelihoods on division by priors, which act as substitutes for the likelihoods of a GMM. The training can be done with/without sequence discriminative criteria mentioned above.

There are several non-HMM based systems based on matching of exemplars or templates of spoken utterance [21] used for spoken term detection and not in the context of LVCSR. However we have not chosen them for this study. Thus, we have used following five AMTs:

1) Monophone based GMM-HMM (MONO)
2) Context-dependent triphone based GMM-HMM (CD-TRI)
3) The Subspace Gaussian models with boosted Maximum Mutual Information (SGMM)
4) DNN with DBN Pretraining (DNN-DP)
5) DNN with state-level MBR (DNN-DP-sMBR)

The first two AMTs are based on generative model of speech traditionally used for speech recognition and use training based on maximum-likelihood (ML) criterion. The latter three are the recently developed AMTs. Among them, the SGMM is based on the sGMM-HMM paradigm but with sequence discriminative training *i.e.,* boosted MMI. DNN-DP and DNN-DP-sMBR fall in DNN-HMM hybrid category. Among the various criteria available for sequence discriminative training in DNN-HMM hybrid category, the state-level MBR (sMBR) is considered in this study as it yields the highest recognition accuracy [22]. Both DNN-DP and DNN-DP-sMBR have common initial stages that involve pretraining of RBMs while the latter has an additional stage for sequence discriminative training based on the state minimum Bayesian risk criterion.

TABLE I
Summary of the codecs used in our study. Except wideband codec, all codecs operate at 8kHz sampling rate.

Codec	Type	BW	BR (kbps)	Source
G.711A	Waveform	Narrow	64	ITU-T [23]
MELP	Parametric	Narrow	2.4	Data Compression[24]
AMR-NB	Hybrid	Narrow	4.40	SoX [25]
AMR-WB	Hybrid	Wide	23.85	3GPP [26]
G.728	Hybrid	Narrow	16	ITU-T [23]
G.729A	Hybrid	Narrow	8	ITU-T [27]
G.729B	Hybrid	Narrow	8	ITU-T [27]
PCM	Waveform	Narrow	128	SoX [25]
ADPCM	Waveform	Wide	32	SoX [25]
GSM-8k	Hybrid	Narrow	13	SoX [25]
SPEEX	Hybrid	Wide	27.8	SPEEX [28]

B. Selection of codecs

The codecs, on the other hand, have been chosen based on the type of codecs (waveform, parametric and hybrid), bandwidth (BW) (narrowband vs wideband) and bit rates (BR) (low vs high). The variety in codec parameters would help in understanding their impact on recognition accuracies *vis-à-vis* the AMTs used. Table I summarizes the details about the codecs considered in the study. It is seen that there are seven hybrid, three waveform and one parametric codecs. In terms of BW, there are three wideband and eight narrowband codecs. The source for the each codec is also indicated in the last column of the table.

Recognition experiments are performed on the speech data obtained after a single round of encoding-decoding. The first eight codecs in Table I are used for this purpose. For the selection of codecs for tandem topologies, we consider various types of distortions. One among these non-linear distortions is due to framewise processing of some codecs (*e.g.,* hybrid type) which have differences in frame size. Additionally, the tandeming of a wideband and narrowband codec would result in loss of information. Such factors could have an impact on the recognition accuracies. In order to model such effects, we have chosen tandem topologies consisting of different combinations of codecs of different bandwidths and frame sizes (last three codecs in Table I). This choice makes all the codecs, used in tandem scenarios, unseen, *i.e.,* none of them were used to train a codec specific acoustic model in the single round of encoding-decoding scenario. This is done to identify the best alternative acoustic model apart from the matched condition models, which is difficult to obtain for a blind tandem topology where the codecs and their orders are unknown.

III. Experimental Details

A. Datasets

Recognition experiments are performed using the TIMIT database [29]. The TIMIT files originally sampled at 16kHz are downsampled to 8kHz depending on the needs of individual codecs. The TIMIT database consists of the training set, complete test set including the core test set [30]. The acoustic models are built from the training set comprising 462 speakers with 3696 utterances. To construct development sets, we use a part of the complete test set, non overlapping with the core test set, comprising 50 speakers with a total of 400 utterances. We create test sets by using the core test set comprising 24 speakers and 192 utterances.

To study the effects of single encoding-decoding, we pass the speech data (wav files) through the encoding and subsequently through decoding function of the speech codecs. We perform this operation on the training, development and test dataset for the first eight codecs from Table I. Hence, we obtain eight codec dependent acoustic models and development and test datasets, each, comprising eight sets of corresponding speech data passed through a single round of encoding-decoding.

In the tandem scenario, to create each tandem topology, the same operation is repeated successively in order of the applicability of codecs for that specific topology. To simulate a blind test in this scenario, we consider the last three codecs from Table I to prepare six blind test databases (3! tandem topologies using 3 codecs) – 1) ADPCM→GSM-8k→SPEEX, 2) ADPCM→SPEEX→GSM-8k, 3) GSM-8k→ADPCM→SPEEX, 4) GSM-8k→SPEEX→ADPCM, 5) SPEEX→ADPCM→GSM-8k, 6) SPEEX→GSM-8k→ADPCM. We also construct a *cocktail* acoustic model by using training data uniformly distributed among these tandem topologies following the work by Srinivasamurthy et al [14], which uses cocktail data for adaptation of the existing models. However, unlike adaptation, we train an acoustic model using cocktail data. This is similar to the work by Sciver et al [31], which does not use data from tandem topologies but a mixed data from several single decodings. It is to be noted that, in both single and tandem scenarios, the effect of the channels is kept out of the scope of this study.

B. Experimental setup

The recognition experiments are carried out using Kaldi toolkit [32] based on the recipes for TIMIT database. In addition to the speech data, the TIMIT corpus contains phonetic transcription corresponding to 61 phones which are compacted into a set of 48 phones during acoustic modeling [30]. We consider a variety of features across different AMTs such as Mel Frequency Cepstral Coefficient (MFCC) with velocity (Δ) and acceleration ($\Delta\Delta$) coefficients, and MFCC with Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT), and Speaker Adaptive Training (SAT) [33]. The Kaldi terminology and features corresponding to the five AMTs are described in Table II. The feature dimensions are 39 for MONO, 40 for CD-TRI and SGMM respectively. The number of probability density function (PDFs) for training of MONO, CD-TRI, SGMM, DNN-DP and DNN-DP-sMBR are 144, 1866, 1973, 1967 and 5681 respectively. Between the two recipes of DNN available in Kaldi-ASR toolkit we choose the Karel's DNN setup for our study as it has been shown to provide better performance [34], [22], [35]. For quantifying recognition performance, we use the word error rate (WER) reported by Kaldi which is referred to as PER in this work.

TABLE II
List of AMTs, Kaldi terminology and their corresponding features.

AMT	Kaldi terminology	Features
MONO	mono	MFCC+Δ+$\Delta\Delta$
CD-TRI	tri3	LDA + MLLT + SAT
SGMM	sgmm2_4 _mmi_b0.1	LDA + MLLT + SAT
DNN-DP	dnn4_pretrain -dbn_dnn	LDA + MLLT + SAT
DNN-DP-sMBR	dnn4_pretrain -dbn_dnn_smbr	LDA + MLLT + SAT

IV. PER analysis for different codecs and AMTs

In this section we elaborate the effects of different codec based acoustic models and AMTs on the PER, under both single encoding-decoding and tandem scenarios.

A. Single encoding-decoding

As mentioned in Section III-A we choose the top eight codecs from Table I to train eight acoustic models under each

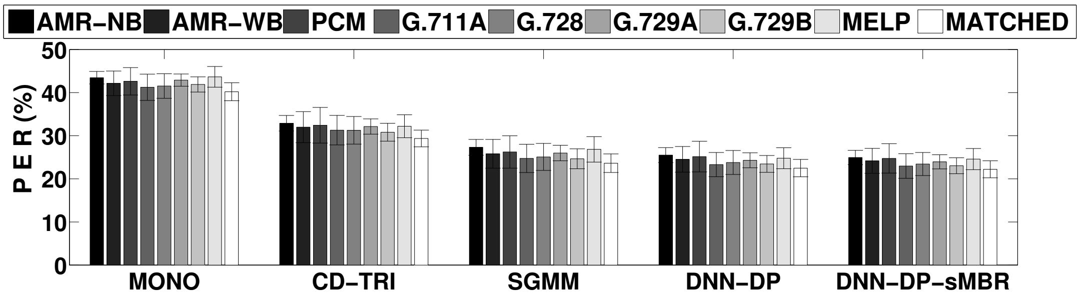

Fig. 1. *The average (standard deviation) PER (%) for all acoustic models and five AMTs across the* ***development sets*** *under single encoding and decoding scenario. Error bars indicate standard deviation.*

AMT. We perform ASR on the development sets for all eight single encoding-decoding. Fig. 1 shows the PER averaged over the development sets by using the eight acoustic models for each of the five AMTs.

It is seen from the Fig.1 that the PER decreases with the improvements in the AMTs from MONO to DNN-DP-sMBR for each codec dependent acoustic model. We see a consistent pattern of PER obtained using eight acoustic models across all the AMTs.

We rank order the eight acoustic models based on the PER for each AMT. We then construct a histogram of the top five acoustic models across AMTs, to find which of them perform consistently well, i.e., low PER values. The histogram of the top five acoustic models across all the AMTs is shown in Fig.2. We find that only six acoustic models based on G.711A, G.728, G.729B, AMR-WB, G.729A and PCM come up as the top five, at least once. We observe that the acoustic models trained on codecs with higher bit rate (Table I) appear in the histogram. This is in agreement with the results reported earlier in the literature. From the Fig. 2, we observe that except for PCM, the other five codecs turn out to be in the top five consistently. Therefore, we use these five codec based acoustic models to perform ASR on the test dataset. Interestingly, we also find that four among the chosen five acoustic models are narrowband codecs.

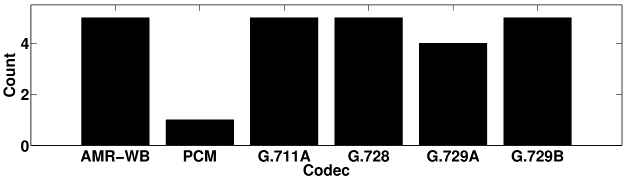

Fig. 2. *Histogram of top four ranked codecs across different AMTs.*

TABLE III
PERCENTAGE CHANGE IN PER RELATIVE TO MATCHED CASE FOR EACH AMT ACROSS DIFFERENT CODECS FOR DEVELOPMENT AND *test* DATASETS UNDER SINGLE ENCODING-DECODING SCENARIO

Trained Codec Model	Data set	Acoustic modeling technique				
		MONO	CD-TRI	SGMM	DNN-DP	DNN-DP-sMBR
AMR-NB	Dev	8.12	12.05	15.61	13.45	12.32
AMR-WB	Dev	4.88	8.94	9.26	9.06	8.84
	Test	*4.38*	*6.98*	*6.90*	*7.56*	*7.80*
PCM	Dev	6.03	10.43	11.06	11.9	11.37
	Test	*5.52*	*10.03*	*10.25*	*10.28*	*10.47*
G.711A	Dev	2.58	6.56	4.76	3.67	3.43
	Test	*2.32*	*6.11*	*5.49*	*2.83*	*3.09*
G.728	Dev	3.30	6.47	6.14	5.78	5.51
	Test	*2.56*	*5.70*	*7.05*	*4.68*	*4.71*
G.729A	Dev	6.72	9.45	9.95	8.06	7.82
	Test	*5.85*	*10.49*	*7.53*	*6.17*	*6.17*
G.729B	Dev	4.23	4.94	4.34	4.34	3.71
	Test	*4.47*	*5.53*	*5.49*	*4.42*	*4.34*
MELP	Dev	8.61	9.66	13.60	10.23	10.69

We compare these PERs with that obtained from the 'matched' condition, *i.e.,* when a codec used to build the acoustic model for recognition is also used in the development set. It is expected that the 'matched' case would yield a lower PER than the other PERs, as shown in Fig.1. To quantify the performance of the eight acoustic models with that of the matched model, we compute the percentage change in PER relative to the matched condition for all AMTs as shown in Table III. We observe that the acoustic models based on G.711A, G.728, G.729B, AMR-WB and G.729A (top five chosen from the histogram as shown in Fig. 2) have less than 10% increase in the PER relative to the matched condition in each AMT. Among the top five, we find that the least increase in PER happens for G.711A based acoustic model, in most of the AMTs. In comparison with the codec based acoustic models, the 'clean' or the PCM based acoustic model performs poorly, with more than 10% increase in the PER relative to the matched condition in most AMTs. This suggests that using a model trained on 'clean' speech data is not suitable to perform ASR on speech degraded due to codec induced distortion. From the table, we find that in four of the five AMTs, the acoustic models based on AMR-NB, PCM and MELP have the highest increase in PER relative to the matched case. With reference to Fig. 2, it is to be noted that these three models did not come up consistently in the top five models chosen based on PER. It could be that on average, the nature of

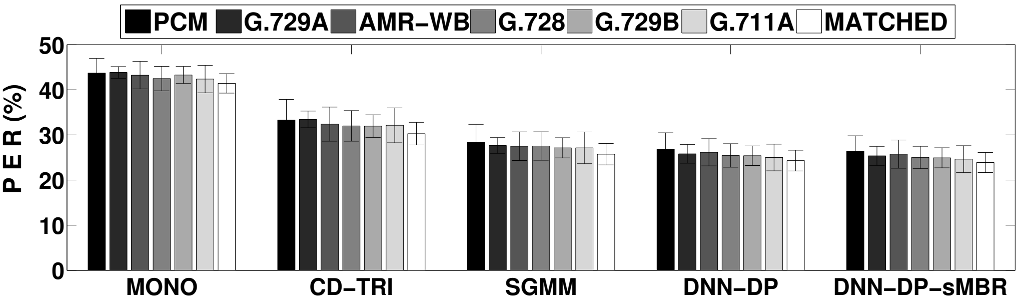

Fig. 3. *The average (standard deviation) PER (%) for all acoustic models and five AMTs across the* ***test sets*** *under single encoding and decoding scenario. Error bars indicate standard deviation.*

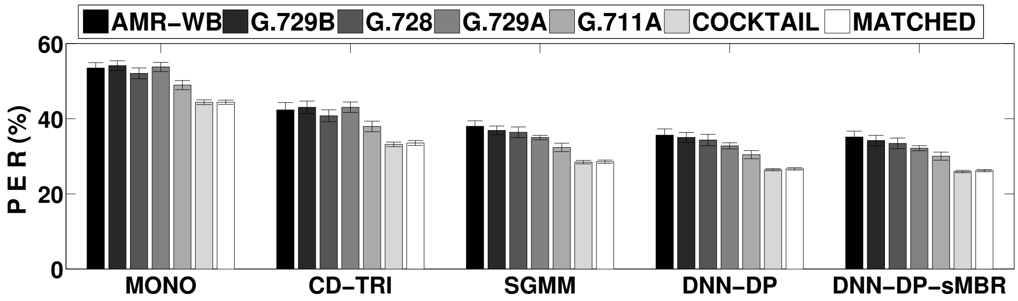

Fig. 4. *The average (standard deviation) PER (%) for all acoustic models and five AMTs across six* ***blind test sets*** *under tandem scenario. Error bars indicate standard deviation.*

distortions[1] induced by these three codecs is different from that induced by the rest of the codecs. This in turn could have led to a greater mismatch between the train and test conditions (on average) yielding a greater deviation from the matched condition (Table III). Interestingly, it is seen from the table that for most acoustic models, the increase in PER relative to the matched case goes up for all AMTs compared to MONO AMT. This suggests that the state-of-the-art AMTs could be less robust to codec induced distortion compared to a traditional GMM-HMM (MONO) AMT.

Fig. 3 shows the average PER obtained by using the top five acoustic models for each of the five AMTs for the test dataset. To compare their performance with the model trained on 'clean' speech data, we also show the PER obtained by the PCM based acoustic model in the figure. As observed for the development dataset, the average PER reduces with the advancements on AMTs (from MONO to DNN-DP-sMBR) on the test dataset as well. From Table III, we see that, for the test set, the top five acoustic models have less than 8% (as opposed to 10% in the case of the development set) increase in the PER relative to the matched condition in most AMTs. Among the top five, we find that the least increase in PER happens for G.711A based acoustic model, similar to the case of the development set. This justifies the choice of using acoustic models trained on narrowband high bit rate codecs to perform ASR of a test data with an unknown coding scheme.

B. Tandem scenarios

We test the performance of the top five codec based acoustic models, obtained from the single encoding-decoding study (Section IV-A), in six blind tandem test scenarios for each AMT. We also compare the performance of the top five acoustic models with that of the *cocktail* acoustic model. Fig. 4 shows the PER obtained by using the top five and the *cocktail* acoustic models averaged over blind test datasets corresponding to six tandem scenarios. The average PER obtained in the matched case is also provided in the figure for each AMT.

Similar to the observations in Section IV-A, we see that the PER reduces with advancements in AMTs. From Fig.4, it is observed that the pattern of PERs from different acoustic model is consistent across all AMTs. Among the top five acoustic models, we see that G.711A has the least PER for each AMT. Incidentally, G.711A has the highest bit rate among the top five codecs chosen (Section IV-A and Table I). Interestingly, we see that the performance of the *cocktail* acoustic model is comparable to that of the matched condition for each AMT. While the *cocktail* acoustic model could be used when the pool of tandem topologies are known, under a blind tandem scenario, acoustic models built on narrowband high bit rate codecs could be used for ASR.

[1] We assume that the distortions in the acoustics introduced are codec specfic. We consider PCM to introduce the least distortion.

V. Conclusions

The present study on the codec induced distortion on the speech recognition performance shows that the acoustic model from G.711A, a narrowband high bit rate codec, results in the best recognition accuracy among acoustic models from eight different codecs using all five types of AMTs considered. This is true for both single round of encoding-decoding as well as blind tandem scenarios. When the pool of tandem topologies are known a priori, *cocktail* acoustic model could be used since it performs better than the acoustic model from G.711A. The study of the effectiveness of the *cocktail* acoustic model in conjunction with the language model, for both single encoding-decoding and tandem scenarios, require further investigation. In addition, to aid the compensation of the codec induced distortions in both these scenarios, acoustic features robust to codec induced distortion could be used for acoustic modeling to improve the recognition performance further. These are parts of our future work.

VI. Acknowledgement

This work was supported by Defence Research and Development Organization, Government of India.

References

[1] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2004.

[2] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow *et al.*, "Subspace Gaussian mixture models for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010*. IEEE, pp. 4330–4333.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[4] S. Euler and J. Zinke, "The influence of speech coding algorithms on automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP '94*, vol. 1. IEEE, pp. 621–624.

[5] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *The 4th International Conference on Spoken Language Processing*, vol. 4. IEEE, 1996, pp. 2344–2347.

[6] H.-G. Hirsch, "The influence of speech coding on recognition performance in telecommunication networks." in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002*.

[7] H. Kook Kim and R. V. Cox, "Bitstream-based feature extraction for wireless speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2000*, vol. 3. IEEE, pp. 1607–1610.

[8] L. V. Grande, I. C. Múgica, L. A. H. Gómez, and E. L. Gonzalo, "Reconocimiento de voz en el entorno de las nuevas redes de comunicacion umts e internet," *Comunicaciones de Telefónica I+ D*, no. 23, pp. 99–112, 2001.

[9] T. Salonidis and V. Digalakis, "Robust speech recognition for multiple topological scenarios of the GSM mobile phone system," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, vol. 1. IEEE, pp. 101–104.

[10] S. Dufour, C. Glorion, and P. Lockwood, "Evaluation of root-normalised front-end (RN LFCC) for speech recognition in wireless GSM network environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96*, vol. 1. IEEE, pp. 77–80.

[11] C. Mokbel, L. Mauuary, D. Jouvet, and J. Monné, "Comparison of several preprocessing techniques for robust speech recognition over both PSN and GSM networks," in *European Signal Processing Conference, 1996*. IEEE, pp. 1–4.

[12] T. Skogstad and T. Svendsen, "Distributed ASR using speech coder data for efficient feature vector representation." in *INTERSPEECH*, 2005, pp. 2861–2864.

[13] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, and K. Bartkova, "Towards improving ASR robustness for PSN and GSM telephone applications," *Speech communication*, vol. 23, no. 1, pp. 141–159, 1997.

[14] N. Srinivasamurthy, S. Narayanan, and A. Ortega, "Use of model transformations for distributed speech recognition," in *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001.

[15] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] V. Valtchev, J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[18] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008*. IEEE, pp. 4057–4060.

[19] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007*, vol. 4. IEEE, pp. 321–324.

[20] M. Gibson and T. Hain, "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition." in *INTERSPEECH*, 2006.

[21] A. Mandal, K. P. Kumar, and P. Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology*, vol. 17, no. 2, pp. 183–198, 2014.

[22] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[23] I. T. Union, *G.191: Software tools for speech and audio coding standardization*, http://www.itu.int/rec/T-REC-G.191-201003-I/en, Last accessed March 27, 2016.

[24] D. Compression, *C source codes for 2.4 kbps MELP coder*, http://www.data-compression.com/melp1.2.tar.gz, Last accessed March 27, 2016.

[25] SoX, *Swiss Army knife of sound processing programs*, http://sox.sourceforge.net/, Last accessed March 27, 2016.

[26] 3rd Generation Partnership Project (3GPP), *ANSI-C code for the Adaptive Multi-Rate - Wideband (AMR-WB) speech codec (Release 13)*, http://www.3gpp.org/ftp/specs/archive/26_series/26.173/, Last accessed March 27, 2016.

[27] I. T. Union, *G.729 : 11.8 kbit/s CS-ACELP speech coding algorithm*, http://www.itu.int/rec/T-REC-G.729-199809-S!AnnE/en, Last accessed March 27, 2016.

[28] Speex, *A Free Codec For Free Speech*, http://www.speex.org/downloads/, Last accessed March 27, 2016.

[29] https://catalog.ldc.upenn.edu/LDC93S1.

[30] C. Lopes and F. Perdigao, "Phone recognition on the TIMIT database," *Speech Technologies/Book*, vol. 1, pp. 285–302, 2011.

[31] J. V. Sciver, J. Z. Ma, F. Vanpoucke, and H. V. hamme, "Investigation of speech recognition over IP channels," in *Proceedings of the IEEE International Conference on Acoustics, Speech,and Signal Processing, ICASSP 2002*, vol. 4. IEEE, pp. 3812–3815.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. of IEEE ASRU*. IEEE Signal Processing Society, 2011.

[33] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *INTERSPEECH*, 2013, pp. 109–113.

[34] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*.

[35] P. Cosi, "A KALDI-DNN-based asr system for Italian," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–5.