Journal of Engineering and Applied Sciences 12 (23): 7293-7298, 2017
ISSN: 1816-949X

Performance Evaluation of Automatic Speech Recognition with Wideband Speech Codecs

[1]D. Nagajyothi and [2]P. Siddaiah
[1]Department of ECE, Vardhaman College of Engineering, Shamshabad, Telangana, India
[2]Department of ECE, University College of Engineering and Technology, Guntur, India

Abstract: WTIMIT which is a derivative of TIMIT emerged as a latest technique for speech quality. The technique has good wideband characteristics over a range of 50-7 kHz. In this study, on the performance of phoneme recognition system has been performed. In the previous paper the effect of decimating the signal to 8 kHz is explained in the conventional case. Further it is possible to extend the evaluation of the AMR-wideband codec for several acoustic models. It is possible to propose the WTIMIT type of wideband channel data from training interactive voice receiving system. The observation is that though some of the codes are showing poor performance at lower bit rates, ASR performance is comparable with other codecs at higher bit rates.

Key words: Speech codecs, IVR, AMR-WB, TIMIT, rates, higher

INTRODUCTION

The typical bandwidth of speech is <4 kHz for applications in telephonic operations. This is often termed as narrow band. The IVR operates at a sampling rate of 8 kHz for Operation (Huang *et al*., 2001; Church and Mercer, 1993; Processing, 2007). Citing this, the advanced speech service system expanded their BW to Wide Band (WB) frequency range of 0.05-7 kHz. The influence of a conventional telephony network on the N-TIMIT is used to evaluate the performance features of recognition system in traditional telephony. The Phoneme Error Rrate (PER) suppressed by a huge extent due to direct WB speech. Similar in NB case, 23% relative PER degradations is identified. It is also reported that there is an evidence of the impact of a WB mobile network using the WTIMIT corpus. An enhancement of 19% PER with respect to direct WB speech is observed while there is a suppressing 3% PER relative to narrow band. In spite of these efforts it is to note that investigation pertaining to effects of telephony network is to be validated. This is more desirous in IVR based telephony system.

The development of DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus paved a way for evaluating Automatic Speech Recognition (ASR) systems (Garofolo *et al*., 1993). It constitutes wideband speech recordings which are sampled at 16 kHz. They typically containing in the rate of 50 Hz to 7 kHz with respect to 630 native speakers. This is with reference to 8 major regions in the US. For training ten phonetically rich sentences are collected from every speech. In every utterance several features are extracted along with speech waveform, time aligned orthographic, phonetic and word transcriptions are taken. With reference to these efforts as of now there are 5 TIMIT derivatives namely FFMTIMIT, NTIMIT, CTIMIT, HTIMIT and STC-TIMIT. The FFMTIMIT can be abbreviated as free field mcrophone TIMIT typical composed of natural TIMIT database. It typically uses a free field device for recording. NTIMIT (Network TIMIT) is adjunct to TIMIT with database constituting the speech wave form (Garofolo *et al*., 1993). Over a telephone handset, similarly CTIMIT constitutes of the original TIMIT recordings were passed through cellular telephone circuits. However, in the case of HTIMIT (Handset TIMIT) the data base consists of two subset with 192 male and female speakers (Garofolo, 1996). The corresponding speech signals are those which are transmitted through different telephone handsets. This typically helps in the investigation of telephone transducer effects on speech. For STCTIMIT which is single channel, the speech signals were sent through a real and in contrast to NTIMIT, all these can be turned as the derivation of wideband speech (ETSI., 1999; Jankowski *et al*., 1990; Morales *et al*., 2008; Reynolds, 1997; Bauer *et al*., 2010). While some are telephony are containing narrowband speech. The sampling is at the

Corresponding Author: D. Nagajyothi, Department of ECE, Vardhaman College of Engineering, Shamshabad, Telangana, India

rate of 8 kHz with a range of 200 Hz to 3.4 kHz. Inspite of all these it is to be noted that there is no availability of real world wideband telephony speech corpus. Several versions of wideband speech codes like G.722 (1988, 1999), G.722.2 (2001) and G.711.1 (2008) have been into operation with several techniques like ADPCM, 3GPP and wide band PCM. It is interesting to note that the wide band telephony speech transmission system is wide available and adaptable. In contrast to ever increasing mobile networks citing this, it its essential to have wideband system in the TIMIT for a wide range of scientific investigations. There are several advantages and applications associated with WBSTS. The integrated speech recognition system provides remote dictation or spelling. This was not a possible case with the earlier telephony system.

In this study, an investigation on the performance of the speech CODECS in terms of bit rates is performed. The analysis is based on the experimentation carried out in MATLAB on windows platform in an i3 with 4 GB RAM and ASR toolkit-SPHINX-3

MATERIALS AND METHODS

Speech codecs: In this study, a brief introduction to the speech codecs is given. The aim of the speech codec is to compress the speech signal in order to reduce the bandwidth and requires minimum storage space. When we will reconstruct, it must be very close to original one. Based on intelligibility and naturalness we will measure perceived quality of the signal. Here, we have considered two types of the networks GSM and VoIP.

Speech recognition setup: The original speech files in the TIMIT database are sampled at 16 kHz. Using this speech database, new speech database for 8 kHz sampling rate is created by down sampling and low pass filtering first. Separate HMMs are created for both the databases with corresponding ASR configuration parameters for testing purpose.

Speech recognition setup for narrowband codecs: All the narrowb and codecs work only with 8 kHz sampling rates. As shown in Fig. 1 when the 16 kHz ASR trained models HMMs are used for narrowband codecs, the original 16 kHz speech data is down sampled to 8 kHz first. This data is encoded and decoded with the respective narrowband codec. The decoded data from the narrowband codecs is up-sampled back to 16 kHz before sending it to the ASR system for recognition analysis as shown in Fig. 2.

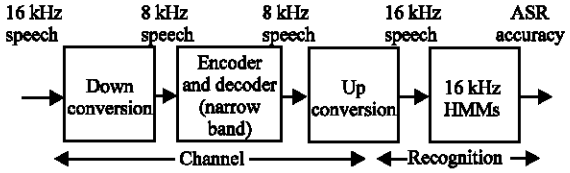

Fig. 1: Recognition with 16 kHz trained models (HMM) for NB codecs

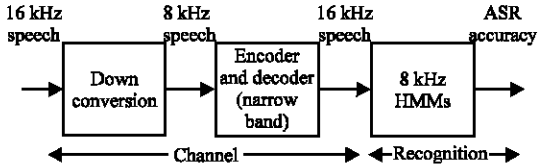

Fig. 2: Recognition with 8 kHz trained models (HMM) for NB codecs

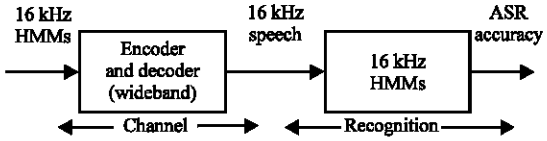

Fig. 3: Recognition with 16 kHz trained models (HMM) for WB codecs

Table 1: ITU-T approved VoIP supported narrowband and wideband speech codecs

Coding standard	Algorithms	Sampling frequency (kHz)	Bit rates (kbps)
G.711 (A/U)	Companded PCM	8	64
G.726	ADPCM	8	16/24/32/40
G.729	CS-ACELP	8	8
G.723.1A	ACELP/MP-MLQ	8	5.3/6.3
G.722 (WB)	SB-ADPCM	16	48, 56, 64
G.711.1 (WB)	Companded PCM, MDCT	16	64, 80, 96
G.729.1 (WB)	CELP, TD-BWE, TDAC	16	8-32
G.722.2 (AMR-WB)	(Multi rate) MRWB-ACELP	16	6.60, 8.85, 12.65,14.25, 15.85,18.25, 19.85,23.05, 23.85

Speech recognition setup for wideband codecs: For wideband codecs analysis with 16 kHz ASR trained models (16 kHz HMMs), sampling conversion is not required as the wideband speech codecs also work with 16 kHz sampling rates. The decoded data from the wideband codecs is directly sent to the ASR system with 16 kHz trained models for recognition as shown in Fig. 3. Incase, 8 kHz trained models are used for wideband codecs, the encoded-decoded speech has to be down-sampled to 8 kHz before applying Table 1 and 2 show the specification summary of the all the supported narrowband wideband codecs.

Table 2: ETSI/3GPP approved GSM supported narrowband and wideband speech codecs

Coding standard	Algorithms	Sampling frequency (kHz)	Bit rates (kbps)
GSM FR	RPE-LTP	8	13.0
GSMEFR	ACELP	8	12.2
GSM HR	VSELP	8	5.6
GSM AMR	(Multi rate) MR-ACELP	8	4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2, 12.2
GSM AMR-WB	MRWB-ACELP	16	6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85

RESULTS AND DISCUSSION

Results pertaining to the technique and proposed method are presented in this section. Testing of the coded data with G.711 coded models (8 kHz HMMs). The 8 kHz coded speech data that is coded with all other wire line codecs such as G.711, G.726 and G.729 is tested with the G.711 coded models (16 kHz trained HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 3.

Case 1: Testing of the coded data with G.711 coded models (16 kHz HMMs): The 8 kHz coded speech data that is coded with all other wireline codecs such as G.711, G.726 and G.729, after up-conversion to 16 kHz is also tested with the G.711 coded models (16 kHz, G.711 coded trained HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 3 and Fig. 4.

Case 2: Testing of the coded data with G.729 coded models (16 kHz HMMs): The 8 kHz coded speech data that is coded with all other wireline codecs such as G.711, G.726 and G.729 is tested with the G.729 coded models (16 kHz HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 4 and Fig. 5.

Case 3: Testing of the coded data with HR coded models (16 kHz HMMs): The 8 kHz coded speech data that is coded with all other NB wireless codecs such as FR, EFR, HR and AMR is tested with the HR-coded models (16 kHz HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 5 and Fig. 6.

Case 4: Testing of the coded data with AMR4.75 coded models (16 kHz HMMs): The 8 kHz coded speech data that is coded with all other NB wireless codecs such as FR, EFR, HR and AMR is also tested with the AMR@4.75 kbps Coded models (16 kHz HMMs)

Table 3: Results of testing the wireline coded data (NB codecs) with G.711coded trained models (16 kHz HMMs)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
G.711	84.78	91.11	92.58	93.75	94.25
G.726	84.70	91.35	92.70	93.86	94.47
G.729	78.55	86.89	89.59	91.34	92.30

Table 4: Results of testing the wireline coded data (NB codecs) with G.729 coded trained models (16 kHz HMMs)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
G.711	85.93	91.14	92.48	93.68	94.14
G.726	85.06	90.67	92.24	93.51	94.18
G.729	84.88	91.18	92.83	93.78	94.47

Table 5: Results of testing the wireless coded data NB codecs with Hr-coded trained models (16 kHz HMMs)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
FR	87.26	92.33	93.60	94.41	94.96
EFR	84.70	91.18	92.92	93.61	94.20
HR	84.72	91.64	92.92	93.91	94.43
AMR@ 4.75	78.41	86.89	89.20	90.63	91.41

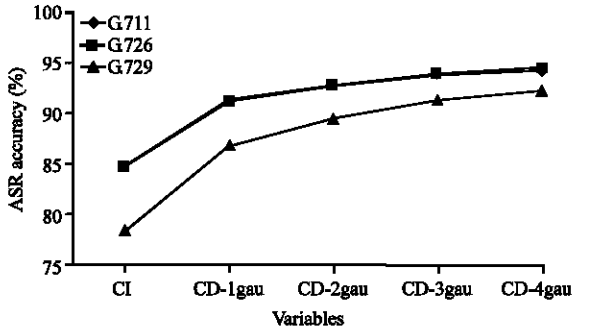

Fig. 4: Graphic results of testing the wireline coded data (NB codecs) with G.711 coded trained models (16 kHz HMMs)

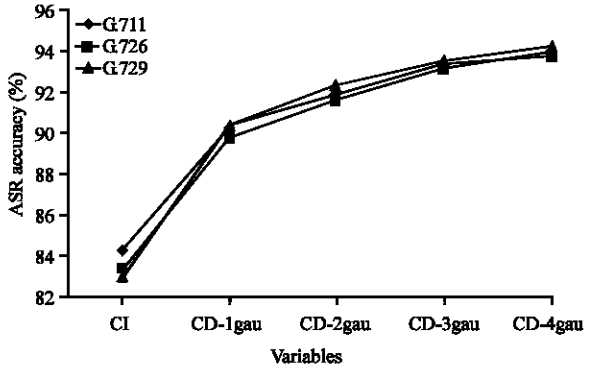

Fig. 5: Graphic results of testing the wireline coded data (NB codecs) with G.729 coded trained models (16 kHz HMMs)

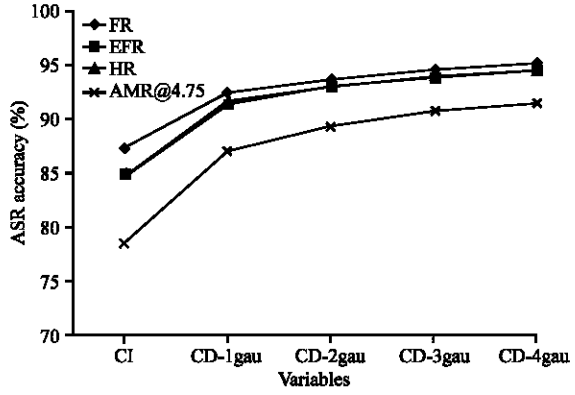

Fig. 6: Graphic results of testing the wireless coded data (NB codecs) with HR-coded trained models (16 kHz HMMs)

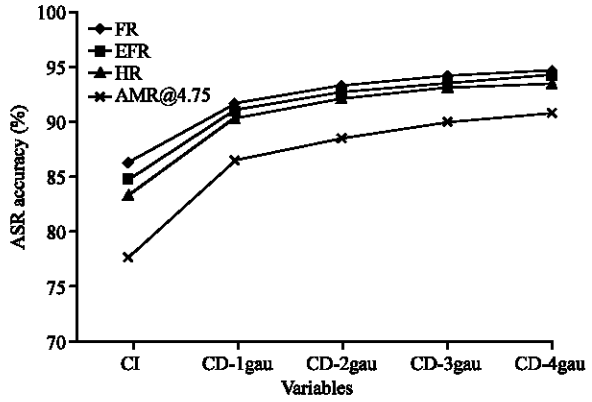

Fig. 7: Graphic results of testing the wireless coded data (NB codecs) with AMR@4.75-coded trained models (16 kHz HMMs)

Table 6: Results of testing the wireless coded data (NB codecs) with AMR@4.75-coded trained models (16 kHz HMMs)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
FR	85.70	91.30	92.53	93.72	94.10
EFR	84.13	90.27	92.15	93.08	93.69
HR	83.15	90.47	91.89	92.98	93.45
AMR@ 4.75	78.61	87.64	90.14	91.75	92.39

for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 6 and Fig. 7).

Case 5: Testing of the AMR coded data with AMR 12.2 Coded Models (16 kHz HMMs): The 8 kHz AMR coded speech data with different bit-rates is tested with the AMR@12.2 kbps Coded models (16 kHz HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 7 and Fig. 8.

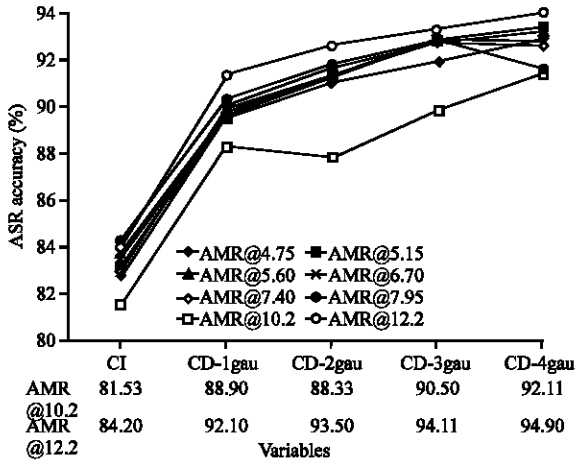

Fig. 8: Graphic results of testing the AMR coded data (NB Codecs) with AMR@12.2 coded trained models (16 kHz HMMs)

Table 7: Results of testing the AMR coded data (NB codecs) with AMR@12.2-coded trained models (16 kHz HMMs)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
AMR@4.75	83.03	90.10	91.70	92.70	93.65
AMR@5.15	83.36	90.23	92.21	93.65	94.25
AMR@5.6	84.11	90.47	92.35	93.47	94.10
AMR@6.7	83.75	90.10	92.10	93.58	93.69
AMR@7.4	84.39	90.70	92.59	93.61	93.45
AMR@7.95	84.57	91.02	92.58	93.67	92.39
AMR@10.2	81.53	88.90	88.33	90.50	92.11
AMR@12.2	84.20	92.10	93.50	94.11	94.90

Table 8: Results of testing the wireless coded data (NB codecs) with FR-coded trained models (16 kHz HMMs)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
FR	86.24	91.68	93.07	94.03	94.48
EFR	84.64	91.04	92.66	93.53	94.34
HR	83.31	90.22	92.04	93.05	93.61
AMR@ 4.75	77.65	86.47	88.45	90.00	90.75

Case 6: Testing of the coded data with FR-coded models (16 kHz HMMs). The 8 kHz coded speech data that is coded with all other NB wireless codecs such as FR, EFR, HR and AMR is tested with the FR-Coded models (16 kHz HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 8 and Fig. 9.

Case 7: Testing of the coded data with EFR-coded models (16 kHz HMMs): The 8 kHz coded speech data that is coded with all other NB wireless codecs such as FR, EFR, HR and AMR is tested with the EFR-coded models (16 kHz HMMs) for the CI and CD-Tied Tri-Phone models with 1, 2, 4 and 8 Gaussians per state and are reported in Table 9 and Fig. 10.

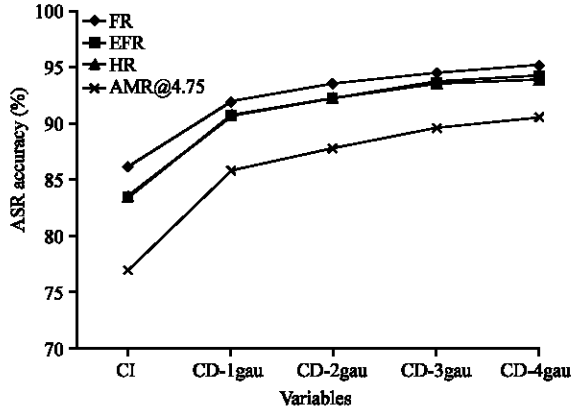

Fig. 9: Graphic results of testing the wireless coded data (NB codecs) with FR-coded trained models (16 kHz HMMs)

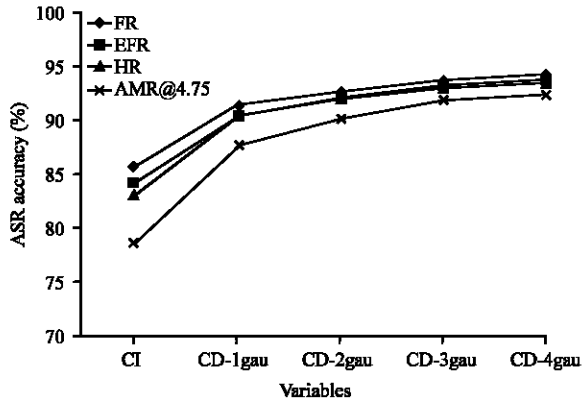

Fig. 10: Graphic results of testing the wireless coded data (NB codecs) with EFR-coded trained models (16 kHz HMMs)

Table 9: Results of testing the wireless coded data (NB codecs) with EFR-coded trained models (16 kHz HMMz)

Coded data used in testing	CI	CD-1gau	CD-2gau	CD-4gau	CD-8gau
FR	86.17	91.88	93.36	94.38	94.93
EFR	83.42	90.56	92.00	93.52	94.03
HR	83.36	90.61	92.11	93.36	93.74
AMR @4.75	76.87	85.67	87.58	89.46	90.43

CONCLUSION

ASR accuracy for un-coded and coded data when tested with the different coded models that include 8 and 16 kHz HMMs is observed. The major observations made are as follows. The ASR accuracy always increases with 8 kHz coded trained models when compared to 8 kHz un-coded models for all the narrowband codecs. The ASR accuracy of coded data of any particular codec increases by at least 2% when the same type of coded models is used. The ASR results for coded data for specific codecs such as G.711, G.729, HR and AMR 12.2 for un-coded and respective coded models are re-organized to see the ASR improvements. Coded data (G.711/G.729/HR/AMR12.2) tested with 8 kHz un-coded HMMs while the coded data (G.711/G.729/HR/AMR12.2) tested with 16 kHz un-coded HMMs. Similarly, the coded data (G.711/G.729/HR/AMR12.2) tested with 8 kHz coded (G.711/G.729/HR/AMR12.2) HMMs whereas the coded data (G.711/G.729/HR/AMR12.2) tested with 16 kHz coded (G.711/G.729/HR/AMR12.2) HMMs. All these codecs perform well for the respective 8 kHz coded models. The ASR performance is almost same when tested with either 16 kHz coded models or 8 kHz un-coded models. The ASR performance is poor when tested with 16 kHz un-coded models.

REFERENCES

Bauer, P., D. Scheler and T. Fingscheidt, 2010. WTIMIT: The TIMIT speech corpus transmitted over the 3G AMR wideband mobile network. Proceedings of the International Conference on Language Resources, Resources and Evaluation, May 17-23, 2010, European Language Resources Association, Valletta, Malta, pp: 1566-1570.

Church, K.W. and R.L. Mercer, 1993. Introduction to the special issue on computational linguistics using large corpora. Comput. Ling., 19: 1-24.

ETSI., 1999. Mandatory speech codec speech processing functions. European Telecommunications Standards Institute, Sophia Antipolis, France.

Garofolo, J.S., 1996. FFMTIMIT. Linguistic Data Consortium, Philadelphia, Pennsylvania, USA. https://catalog.ldc.upenn.edu/LDC96S32.

Garofolo, J.S., L.F. Lamel, W.M. Fisher, J.G. Fiscus and D.S. Pallett *et al.*, 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium Publication, Philadelphia, Pennsylvania, USA. isBN: 1-58563-019-5.

Huang, X., A. Acero and H.W. Hon, 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, Upper Saddle River, New Jersey, USA. isBN:9780130226167, Pages: 980.

Jankowski, C., A. Kalyanswamy, S. Basson and J. Spitz, 1990. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP-90, April 3-6, 1990, IEEE, Albuquerque, New Mexico, pp: 109-112.

Morales, N., J. Tejedor, J. Garrido, J. Colas and D.T. Toledano, 2008. STC-TIMIT: Generation of a single-channel telephone corpus. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08), May 28-30, 2008, European Language Resources Association, Paris, France isBN:2-9517408-4-0, pp: 391-395.

Processing, S., 2007. Transmission and quality aspects (STQ): Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ETSI, Sophia Antipolis, France.

Reynolds, D.A., 1997. HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97 Vol. 2, April 21-24, 1997, IEEE, Munich, Germany is BN:0-8186-7919-0, pp: 1535-1538.