THE INFLUENCE OF SPEECH CODING ON RECOGNITION PERFORMANCE IN TELECOMMUNICATION NETWORKS

Hans-Günter Hirsch

Niederrhein University of Applied Sciences
hans-guenter.hirsch@hs-niederrhein.de
http://dnt.kr.hs-niederrhein.de

Abstract

The influence of encoding and decoding speech on automatic speech recognition is investigated in this paper with respect to applications in today's telecommunication networks. The deterioration of recognition performance is presented for several coding schemes in GSM and future mobile networks. The extraction of acoustic features for the recognition is done with the already standardized ETSI frontend and with the advanced robust frontend whose standardization is almost finished. The Aurora2 experiment for recognizing the noisy TIDigits is taken as experimental basis. Finally recognition results are compared to results of subjective listening tests that have been performed for the characterisation of these speech coding schemes.

1. INTRODUCTION

Applying speech services based on automatic speech recognition in today's telecommunication networks has to take into consideration the influence of a big variety of different speech en(de)coding techniques. Several coding schemes are especially applied in mobile communication for transmitting speech over the bandlimited cellular channel.

Investigations have been carried out to determine the influence of speech coding on the performance of speech recognition systems [1,2] and to improve recognition performance in such situations [3,4,5]. Most of this work has a focus on the GSM full-rate coding scheme that was introduced as first coding technique in GSM mobile networks. In the meantime several other coding schemes are or will soon get available in GSM networks and in mobile networks of the next generation. With the introduction of the AMR (Adaptive Multi-Rate) coding technique a complete set of 8 schemes will be introduced with data rates between 4.75 kBit/s and 12.2 kBit/s. The influence of the different coding techniques is presented in this paper without considering the additional influence of transmitting the coded speech over an erroneous cellular channel. The speech recognition is based on two techniques for extracting the acoustic features. The first approach has already been standardized by ETSI [6]. The second scheme will be adopted as another ETSI standard in the near future [7]. The recognition is based on HMMs by using the training and recognition modules of the HTK toolkit [8]. Results are presented for the task of recognizing noisy digits. The recognition scheme as well as the database have been created in the Aurora group [9].

The influence is shown when training the recognition system on data processed with one of the speech coding schemes and recognizing data en(de)coded with a different technique.

2. EXPERIMENTAL SETUP

Figure 1 gives an overview about the whole experimental setup. Different speech coding schemes are investigated as used in fixed telecommunication networks and in the GSM as well as in future mobile networks. All techniques are listed in table 1 together with their data rates and abbreviations as used throughout this paper.

Coding scheme	Data rate/ kBit/s	Abbreviation
No coding		PCM
G.711 (alaw)	64	ALAW
GSM full-rate	13	FR
GSM half-rate	5.6	HR
GSM enhanced full-rate	13	EFR
AMR (Adaptive Multi-Rate)	4.75	AMR475
	5.15	AMR515
	5.9	AMR59
	6.7	AMR67
	7.4	AMR74
	7.95	AMR795
	10.2	AMR102
	12.2	AMR122

Table 1: List of investigated coding schemes

The processing is done by applying the software as available from ITU and ETSI. The possibility of detecting speech pauses is available as part of most of the coding schemes as so called DTX mode to enable a discontinuous transmission. This mode is not used in all experiments presented in this paper.

ICSLP2002 (7th International Conference on Spoken Language Processing)
Denver, USA, 16-20 September 2002

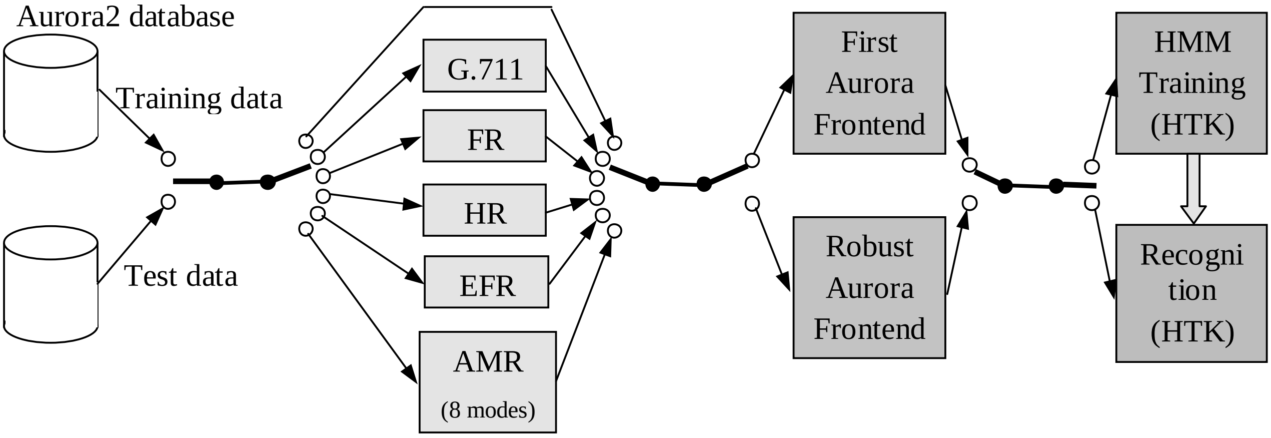

Figure 1: The experimental setup

The feature extraction can be done with the already standardized Aurora frontend [6] or with an advanced noise robust frontend [7] that will be standardized in the near future. The first frontend is more or less an usual cepstral analysis scheme where 13 Mel frequency cepstral coefficients (MFCCs) are calculated and the frame energy as a fourteenth parameter. The advanced frontend contains additional processing blocks for reducing the influence of background noise and compensating the bad effects of an unknown frequency characteristic e.g. caused by the microphone. The output consists also of 13 cepstral coefficients and the frame energy. The advanced frontend contains also a VAD to enable a DTX mode where the acoustic parameters might not be transmitted or not evaluated during speech pauses. As with the speech coding algorithms this DTX mode is not applied for the investigations presented in this paper. Both frontends contain additional compensation and coding schemes to transmit the acoustic features as a data stream at 4.8 kBit/s. It turned out that this compensation has almost no influence on the recognition performance. It is not applied in case of the already standardized frontend but is used with the advanced frontend.
Both processing schemes are designed to handle speech data sampled at 8, 11 or 16 kHz. Only the 8 kHz mode is applied for these investigations because of the limitation to 4 kHz bandwidth due to the coding schemes.

Recognition experiments have been run with the noisy TIDigits database also referred to as Aurora-2 database. Furthermore a HTK based HMM recognizer has been applied as used in the Aurora evaluation process [9]. The digits are modeled as whole word HMMs with the following parameters:

- 16 states per word (according to 18 states in HTK notation with 2 dummy states at beginning and end)
- simple left-to-right models without skips over states
- mixture of 3 Gaussians per state
- only the variances of all acoustic coefficients (No full covariance matrix)

In case of the first frontend 12 MFCCs and the frame energy are taken as acoustic parameters. A single feature vector consists of 39 components by adding the delta and acceleration coefficients as defined inside HTK. For the second frontend the zeroth cepstral coefficient and the frame energy are combined as a single component of the feature vector. Delta and acceleration coefficients are calculated with a slightly different approach as defined in the standard leading also to feature vectors with 39 components in total.

3. RECOGNITION RESULTS

3.1. First ETSI frontend

Recognition results are listed in table 2 when processing all training and test data with one of the coding schemes and extracting acoustic features with the already standardized frontend [6].

Coding scheme	Total word accuracy(%)	Word accuracy (%) (multi-condition only)
PCM	73.23	86.39
ALAW	70.15	85.76
FR	68.31	85.28
HR	66.44	82.45
EFR	71.44	86.22
AMR475	70.16	84.89
AMR515	71.17	84.49
AMR59	69.46	85.05
AMR74	67.58	85.74
AMR102	68.38	85.63

Table 2: Word accuracy for training and testing the recognizer in the same coding mode

The numbers in column 2 of table 2 describe the total word accuracy as an weighted average over 6 individual experiments where each experiment again considers several different noise conditions at a range of SNRs between 0 and 20 dB. All results are worse in comparison to the result without additional speech coding (PCM). As expected the additional encoding and decoding leads to a deterioration of the recognition performance. The lowest accuracy is achieved for the half-rate scheme with a loss of about 7% word accuracy.

3 of the 6 experiments are based on training with clean data only. The average word accuracy for these 3 experiments is in the range of 50 to 60%. Looking at such a low accuracy leads to a more random behavior with higher confidence intervals. This might be one reason for a certain inconsistency of the total results, e.g. the lower accuracy for AMR at 7.4 or 10.2 kBit/s in comparison to the 4.75 kBit/s mode.

The other 3 experiments are based on training with clean and with noisy data in almost the same range of SNRs, also referred as multi-condition training mode. The average results for these 3 experiments only are listed in the third column. These results show a more consistent behavior (as expected) e.g. with an better recognition performance for AMR modes with increasing data rate in almost all cases.

Recognition results are listed in table 3 for a few combinations of training the recognizer in one selected mode and testing with data processed in other coding modes. Best performance is achieved on average when training the recognizer on data without additional speech coding. This might be important to know for applying speech services in telecom networks with access from fixed and mobile networks. But it has to be taken into account that the influence of the cellular channel is not considered here.

3.2. Advanced noise robust frontend

Recognition results are presented in table 4 when processing all training and test data with one of the coding schemes and extracting acoustic features with the advanced robust frontend [7]. Word accuracy is again shown as average value over all 6 experiments and as average over the 3 experiments in multi-condition training mode. The average total accuracy is about 15 to 16 % higher in comparison to the first standardized ETSI frontend. This shows impressively the introduced robustness of the advanced feature extraction scheme especially when training on clean data only.

The recognition is again worse in all cases of applying speech encoding and decoding. The worst result is again achieved for the half-rate coding scheme. The total performance is about 7% less in terms of word accuracy. In comparison to the first ETSI frontend with also 7% loss for the half-rate scheme this corresponds too a much higher relative degradation. Fairly consistent results can be observed when looking at the increase of word accuracy for the increasing data rates of the AMR modes. This holds true for the total results as well as for the average results in case of training on multi-condition data.

Coding scheme	Total word accuracy(%)	Word accuracy (%) (multi-condition only)
PCM	89.30	91.55
ALAW	88.88	91.53
FR	87.31	90.28
HR	81.77	87.18
EFR	88.16	90.97
AMR475	84.76	88.99
AMR515	84.23	89.09
AMR59	85.02	89.66
AMR67	84.90	89.85
AMR74	85.23	90.01
AMR795	85.36	89.72
AMR102	86.36	90.5
AMR122	87.07	90.8

Table 4: Word accuracy for training and testing the recognizer in the same coding mode

Table 5 contains all results when training the recognizer on data without additional coding or training on data in one of the AMR coding modes and testing in all AMR modes. Again best performance is achieved on average when training the recognizer on speech data without additional en(de)coding. The average word accuracy increases when taking training data of an AMR mode at higher data rate.

Another experiment is run by taking the HMMS as determined from the AMR122 mode and applying a few further iterations of embedded Baum-Welch reestimation with all training data at all AMR modes. This leads to HMMs trained on data of all AMR modes. Results are shown in table 6. No real improvement can be observed. It seems to be best again training the recognizer on data with highest speech quality.

4. COMPARISON TO SUBJECTIVE LISTENING RESULTS

The quality of speech encoding and decoding schemes is determined with subjective listening tests. Speech data are processed with the different coding techniques in several languages. The data consist of clean speech as well as of speech with additional background noise. In case of mobile communication a simulation of the transmission over the

Coding scheme applied to training data	Total word accuracy (%) for applying different coding schemes to test data							
	PCM	ALAW	FR	HR	EFR	AMR475	AMR74	AMR102
PCM	73.23	73.44	73.04	67.02	71.93	71.83	71.92	72.14
EFR	68.68	69.10	69.79	66.59	71.44	73.04	72.21	71.54
AMR475	63.53	64.10	65.38	63.18	67.09	70.16	68.51	67.20

Table 3: Word accuracy for training and testing in different coding modes with the first ETSI frontend

cellular channel can also be applied to these data. The processed speech signals are presented to listeners that rate their subjective impression on a scale between 1 and 5 with 5 being the best. Averaging these subjective ratings over a big set of speakers and several languages and different noise conditions leads to so called MOS (mean opinion score) values. Such MOS values are listed in table 7 together with the recognition results (as already shown in table 4) when applying the advanced frontend and looking at the multi-condition training mode. The MOS values are taken from [10] and have been derived from listening to clean speech. A fairly good correlation can be seen between the subjective listening results and the objective recognition results. There might be a good chance to use such recognition experiments as an additional and cost saving method for the objective evaluation of speech coding algorithms. It might be worth to investigate other and more complex recognition tasks in this respect.

5. REFERENCES

[1] S. Euler, J. Zinke: ”The Influence of Speech Coding Algorithms on Automatic Speech Recognition“, ICASSP98, Adelaide, Australia, pp. 621-624, 1994.

[2] P. Haavisto: “Speech Recognition for Mobile Communications”, COST workshop on robust methods for speech recognition in adverse conditions, Tampere, Finland, 1999.

[3] L. Fissore, F. Ravera, C. Vair: “Speech Recognition over GSM: Specific Features and Performance Evaluation”, COST workshop on robust methods for speech recognition in adverse conditions, Tampere, Finland, 1999.

[4] C. Mokbel et al.: “Towards Improving ASR Robustness for PSN and GSM Telephone Applications”, Speech Communication, Vol.23, pp.141-159, 1997.

[5] J.M. Huerta, R.M. Stern: “Distortion-class Modeling for Robust Speech Recognition under GSM RPE-LTP Coding”, Speech Communication, Vol.34, April 2001.

[6] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108 v1.1.2 (2000-04), Apr. 2000.

[7] ETSI draft standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 202 050 v0.1.0 (2002-04), Apr. 2002.

[8] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, “The HTK book – version2.2”, Entropic, 1999.

[9] H.G. Hirsch, D. Pearce: “The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions”, ISCA workshop ASR2000, Paris, France, Sep. 2000.

[10] 3GPP technical report: “Performance Characterization of the AMR Speech Codec”, 3GPP TR 26.975, Jan. 2001

Coding scheme applied to training data	Total word accuracy (%) for applying different coding schemes to test data						
	AMR475	AMR515	AMR59	AMR67	AMR74	AMR795	AMR102