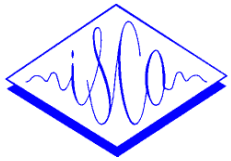
ISCA Archive
http://www.isca-speech.org/archive



7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

NETWORK-BASED VS. DISTRIBUTED SPEECH RECOGNITION IN ADAPTIVE MULTI-RATE WIRELESS SYSTEMS

Tim Fingscheidt, Stefanie Aalburg, Sorel Stan, and Christophe Beaugeant

Siemens AG, ICM Mobile Phones, Grillparzerstrasse 10-18, 81675 Munich, Germany
{first name}.{last name}@siemens.com

ABSTRACT

Distributed speech recognition (DSR) is motivated by the fact that codecs used in speech transmission usually reveal a degrading voice quality below some channel quality (carrier-to-interferer ratio C/I), which justifies efficient coding of features with an appropriate channel coding in the mobile terminal. The Adaptive Multi-Rate (AMR) speech codec standardized for GSM and UMTS however delivers an acceptable speech quality way down to C/I ratios of about 4 dB in the GSM full-rate speech channel.

In this paper we investigate *network-based* speech recognition (NSR) using a conventional speech channel with AMR coding as an alternative to a DSR system. This approach is natural and attractive since information services usually require a duplex channel for conversation anyway, furthermore no change to existing mobiles is required. For a GSM full-rate channel it turns out that an NSR system based on AMR coding indeed is comparable to DSR approaches.

1. INTRODUCTION

1.1. Basic Approaches

Services based on speech recognition used in mobile networks such as GSM and UMTS can be implemented in a variety of conceptual solutions. A first approach could be to perform the full speech recognition solution in the mobile terminal itself. Having hardware constraints and power consumption in mind, such *terminal-based* (TSR) solutions allow however only for very restricted functionality and are therefore not regarded further in our investigations. The classical solution which is widely employed today is the *network-based* speech recognition (NSR). NSR how we understand it uses a full-duplex speech channel with speech coding (for bit rate reduction) and channel coding (for error protection/correction). A third approach is to avoid the speech coding in the uplink channel (connection from talker to speech recognition system in the network). Instead, the mobile terminal only extracts and codes the feature vectors (front-end), the transmission is performed over a data channel, and the recognition engine is located on the server side (back-end). ETSI has followed this so-called *distributed* speech recognition (DSR) approach for some time already in the STQ Aurora DSR working group. Performing speech recognition in the network (as in NSR and DSR) provides a number of significant advantages for wireless systems. The most important one is probably the computational burden which is shifted from the mobile terminal to a high performance platform in the network.

It turns out that DSR provides a good coding efficiency. This has been shown by the AURORA work which resulted in a front-end yielding a bit rate of 4.4 kbps for the compressed feature vectors. Including header information and CRC bits for error detection the bit rate increases to 4.8 kbps allowing to use an appropriate GSM full-rate *data* channel with a gross bit rate of 22.8 kbps [1]. An interesting approach to DSR is to interprete the speech encoder as front-end and to run the back-end on the speech codec parameters as features (see e.g. [2],[3]). This has been shown to provide promising results.

Much of the present DSR work is based on the assumption that a comparable NSR approach uses a fixed bit rate speech coder, e.g. the GSM Enhanced Full-Rate (EFR) coder [4] running at 12.2 kbps in the GSM full-rate *speech* channel with a gross bit rate of 22.8 kbps (see e.g. Fig. 4 in [5]). In the meantime however new speech services have emerged such as the AMR speech coder [6] which is an optional coder for GSM and the mandatory speech coder for UMTS networks.

1.2. AMR Coding in NSR

The AMR coder operates in 8 different coding rates (modes), namely at 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15, and 4.75 kbps, respectively. The modes can be switched seamlessly to guarantee an optimal sharing of the gross bit rate of the speech coder and the channel coder de-

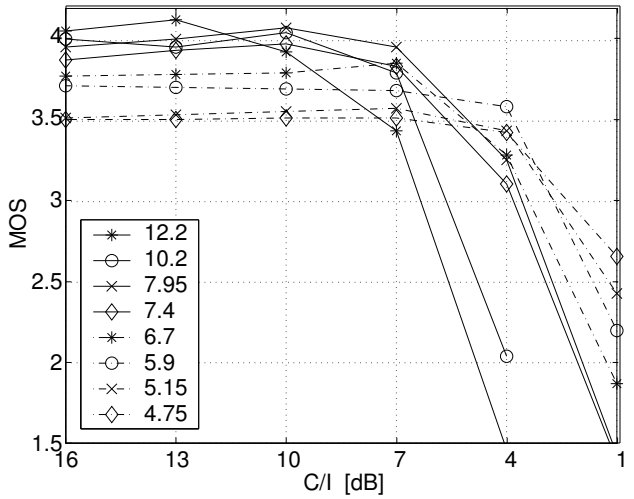


Figure 1: MOS curves for AMR coded speech over a GSM full-rate speech channel: American English and German. The MOS values are taken from [7].

pendent on the quality of the transmission link. In low C/I cases (bad channel conditions) low bit rate modes are chosen in order to allow for a strong error protection. In good channel conditions the higher AMR modes are chosen to deliver the best speech quality available.

In Fig. 1 the *mean opinion score* (MOS) curves measured for all 8 modes for C/I ratios from 16 dB (very good channel) down to 1 dB (extremely bad channel) are plotted [7]. It can easily be seen that there is a MOS difference of about 0.5 points in error free channel conditions. Moreover, all AMR modes maintain their speech quality down to C/I = 10 dB. The two highest bit rate modes show some performance degradation at 7 dB C/I, the next three modes degrade at 4 dB C/I, while the modes with 5.9 kbps and lower bit rates keep their quality over a wide range of channel conditions and finally degrade at 1 dB C/I. Formulated vice versa: For a C/I ratio of 1 dB and below the optimum bit rate *in terms of speech quality* is 4.75 kbps, for 4 dB the 5.9 kbps mode is best, for 7 and 10 dB 7.95 kbps is optimum and finally for higher C/I ratios 12.2 kbps should be chosen. If switched at the right points, the adaptive multi-rate system yields the envelope curve of the depicted modes. The switching strategy for AMR speech transmission is given in [8]. It should be noted that no significant language dependency of the AMR speech codec was observed in terms of MOS.

While earlier publications on NSR mostly focused on fixed-bit rate speech codecs, this paper benchmarks the performance of the NSR approach using the AMR

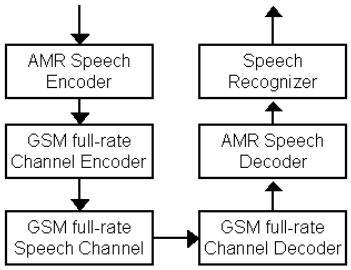


Figure 2: Simulation chain for the NSR system.

speech coder with its powerful mode switching strategy. The results we obtain may be interpreted as the *machine* complement to Fig. 1, which is based on speech quality perception by *human beings*. Finally, a comparison is done to the classical DSR approach.

2. THE SPEECH RECOGNIZER

The speech recognition engine used in our simulations is a recognizer actually optimized for embedded systems with low memory footprint. It uses continuous density HMMs with unified variance in Bakis topology. More details on the recognition itself can be found in [9]. For the digits task used in our simulations we employed whole word HMM models for the digits based on a total of 1200 prototypes.

The front-end works on 8 kHz sampled speech. Preemphasis, Hamming windowing, and a 256 sample FFT are performed. The frame shift yields one frame each 15 ms. In the frequency domain a spectral subtraction for noise reduction is performed. Subsequently, an energy value as well as 12 MFCC coefficients are computed, where the latter are subject to a cepstral liftering technique. Finally, delta and acceleration coefficients are computed. The resulting 39 values are concatenated with those from the adjacent frame yielding 78 values subject to an LDA transformation. Only the 24 most important LDA features constitute the actual feature vector which is available then every 15 ms. Some additional information on the underlying front-end can be found in [10].

3. SIMULATION RESULTS

As depicted in Fig. 2 the simulation chain consists of an AMR speech encoder, a GSM full-rate channel encoder, an error insertion device simulating the GSM full-rate channel, an appropriate channel decoder, an AMR decoder, and finally our speech recognition system with front-end and back-end. The channel reflects a typical urban profile with a user moving at 3 km/h and ideal frequency hopping.

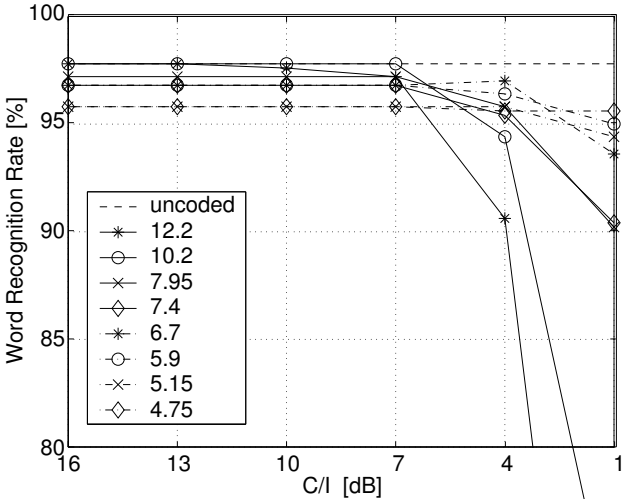


Figure 3: Word recognition rates (WRR) in [%] for AMR coded speech over a GSM full-rate speech channel: German isolated digits.

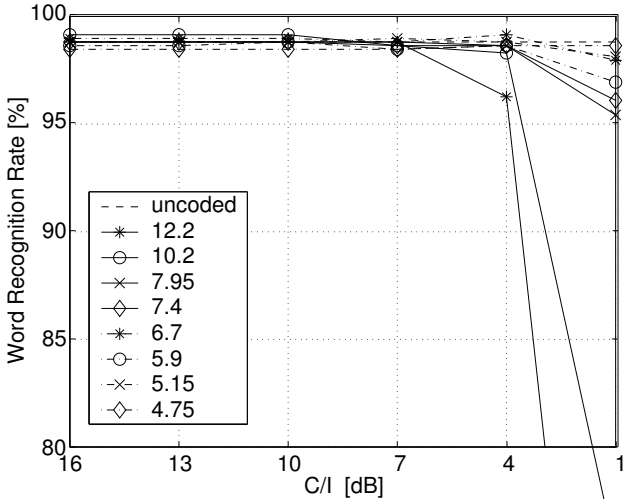


Figure 4: Word recognition rates (WRR) in [%] for AMR coded speech over a GSM full-rate speech channel: Spanish isolated digits.

The task consists of single digits in German (501 samples) and Spanish (590 samples). They are part of the SpeechDat-Car database recorded in a car environment using a close microphone (see e.g. [11]). The location of a noise reduction scheme (either in front of the speech encoder in the mobile or integrated into the network-based feature extraction) is of high importance in an NSR approach. Moreover, it turns out that the optimum parameterization of the noise reduction is quite dependent on its location. Since we are not mainly interested in noise reduction issues, we chose digit samples reflecting low car speeds and town traffic, thus ensuring that the location and the parameterization of the noise reduction are not that critical. In an optimum system design however, some kind of noise reduction should be carried out in front of the speech encoding already in order to prevent the encoder from a speech production model mismatch.

In Fig. 3 the results for the German digits are shown. Direct recognition on the given database yields a word error rate (WER) of 2.2 %. This serves as a baseline in the following discussion. A DSR system with the same front-end as used in our network-based recognizer would yield the same or lower (due to quantization of features) quality in channel error free conditions. Note that any better front-end improves both the performance of a DSR and the NSR system.

The same value of 2.2 % is achieved with AMR coding at the bit rates 12.2 and 10.2 kbps. Fig. 3 shows that for German digits the NSR system is comparable in performance with a DSR system at C/I ratios of 7 dB and higher.

It could be argued however, that a bit rate of 10.2 kbps is much higher than a DSR approach would require (e.g. the AURORA front-end yielding 4.8 kbps). Surely, however the gross bit rates of both systems in GSM are equal, i.e. the GSM full-rate channel operating with 22.8 kbps is used in both cases. The only difference is that a data channel is used for DSR, while the NSR scheme runs on the speech channel. Accordingly, the same bandwidth resources are required in both cases.

For C/I ratios lower than 7 dB the AMR-based NSR system now chooses lower bit rate modes down to 4.75 kbps at a C/I of 1 dB. Table 1 shows the best modes for a given C/I ratio with the respective WER of the NSR system.

Figure 4 gives the results for the Spanish digits task. Besides a better baseline WER of 1.2% we encounter an improvement by speech coding: The WERs of the best AMR modes down to C/I = 4 dB are better than the uncoded baseline performance, and thus exceed the performance of any potential DSR system based on our front-end. The reason for such a behaviuor can be the spectral whitening effect to background noise in the presence of speech when running through a speech transcoder.

This behaviour was not found for German, however. In total it can be summarized that down to about 4...7 dB the system behaviour of the proposed NSR scheme is found being comparable to a DSR scheme.

For C/I = 1 dB still WERs of 4.4% (German) and

C/I		–	16	13	10	7	4	1
DE	**WER**	2.2	2.2	2.2	2.2	2.2	3.0	4.4
	Mode	–	10.2	10.2	10.2	10.2	6.7	4.75
ES	**WER**	1.2	0.8	0.8	0.8	1.0	0.8	1.4
	Mode	–	10.2	10.2	10.2	5.15	6.7	4.75

Table 1: Word error rates (WER) [%] of the best AMR mode for a given C/I ratio in dB. DE: German, ES: Spanish. "–" means no AMR coding and transmission at all, but direct recognition on the speech database.

1.4% (Spanish) can be achieved, which is acceptable for a situation, where the mobile might even loose already its synchronization to the network. Note that in the C/I = 1 dB case even the net bit rates of our NSR scheme and a DSR scheme (e.g. AURORA) are comparable.

Looking at Table 1 it turns out that the experiments done in German and Spanish show a quite consistent picture in terms of AMR mode selection based in C/I measurements. Only C/I = 7 dB yields 10.2 kbps mode in one case and 5.15 kbps mode in the other case. For Spanish this mode however is somewhat surprising (since it is lower than the optimum mode for 4 dB), thus the best 7 dB mode should be somewhere between 6.7 and 10.2 kbps.

Having proven the surprisingly good performance of the proposed NSR system, it should be noted that no additional latency is required beyond that encountered in a speech channel for normal conversation. This is an advantage compared to most known DSR approaches.

In the same way as DSR, NSR moves the computational load from the mobile terminal to the network. NSR even goes further: No additional computations are needed (and stored as program code!) except those which are used in speech conversation anyway. The only change to the AMR system for conversational speech might be a different link adaptation scheme (selection of best modes dependent on C/I measurements) [8]. Comparing Figs. 3 and 4 to Fig. 1, it is an interesting observation that the speech recognizer obviously is more robust to channel errors as the human ear, since in the MOS figure significant degradations start at somewhat higher C/I ratios as in the WRR figures. This leads to the conclusion that MOS figures do not really provide information on recognition performance.

4. CONCLUSIONS

We propose a network-based speech recognition (NSR) system using the AMR speech coder for transmission. Simulation results of a German and Spanish digits task using the GSM full-rate channel and our speech recognition system are presented. It turns out that down to C/I ratios of about 4...7 dB the proposed NSR system achieves a maximum quality comparable to that of an error free, uncoded speech transmission, and is therefore also comparable to a DSR approach. The achieved results encourage the use of the NSR approach in wireless systems employing the adaptive multi-rate (AMR) speech coder. An attractive side effect is that no changes to the mobile terminals are necessary.

REFERENCES

[1] "ETSI STQ Aspects: Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms," ETSI ES 201 108 V1.1.2, Apr. 2000. 1

[2] H.K. Kim and R.V. Cox, "A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 558–568, July 2001. 1

[3] B. Raj, J. Migdal, and R. Singh, "Distributed Speech Recognition with Codec Parameters," in *Proc. of ASRU'01*, Madonna di Campiglio, Italy, Dec. 2001. 1

[4] "Digital Cellular Telecommunications System (Phase 2+); Enhanced Full Rate (EFR) Speech Transcoding," ETSI EN 300 726 V8.0.1, Nov. 2000. 1

[5] D. Pearce, "An Overview of the ETSI Standards Activities for Distributed Speech Recognition," in *Proc. of AVIOS'00*, San Jose, CA, USA, May 2000. 1

[6] "3GPP TSG SA: Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; Transcoding Functions," 3G TS 26.090 V3.1.0, Dec. 1999. 1

[7] "3GPP TSG SA4: Codec; Performance Characterization of the AMR Speech Codec," 3G TS 26.075 V1.0.0, Dec. 1999. 2

[8] "3GPP; TSG GSM/EDGE; Radio Access Network; Link Adaptation," 3GPP TS 05.09 V8.4.0, Aug. 2001. 2, 4

[9] J.G. Bauer, "Enhanced Control and Estimation of Parameters for a Telephone Based Isolated Digit Recognizer," in *Proc. of ICASSP'97*, München, Germany, Apr. 1997, vol. 2, pp. 1531–1534. 2

[10] A. Hauenstein and E. Marschall, "Methods for Improved Speech Recognition over Telephone Lines," in *Proc. of ICASSP'95*, Detroit, Michigan, May 1995, pp. 425–428. 2

[11] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "SPEECHDAT-CAR. A Large Speech Database for Automotive Environments," in *Proc. of LREC'00*, Athens, Greece, June 2000. 3