A COMPARISON OF SOFT-FEATURE DISTRIBUTED SPEECH RECOGNITION WITH CANDIDATE CODECS FOR SPEECH ENABLED MOBILE SERVICES

Valentin Ion, Reinhold Haeb-Umbach

University of Paderborn
Dept. of Communications Engineering
33098 Paderborn, Germany
{ion,haeb}@nt.uni-paderborn.de

ABSTRACT

In this paper we present a comparison of the recently proposed Soft-Feature Distributed Speech Recognition (SFDSR) with the two evaluated candidate codecs for Speech Enabled Services over wireless networks: Adaptive Multirate Codec (AMR) and the ETSI Extended Advanced Front-End for Distributed Speech Recognition (XAFE). It is shown that SFDSR achieves the best recognition performance on a simulated GSM transmission, followed by XAFE and AMR. We also present some new results concerning SFDSR which demonstrate the versatility of the approach. Further, a simple method is introduced which considerably reduces the computational effort.

1. INTRODUCTION

Speech Enabled Services (SES) over wireless networks have gained considerable interest due to the potentially high commercial impact of offering such services to the huge number of mobile phone subscribers [1]. Basically two options for the speech codecs have been considered:

- AMR and AMR-WB, which are the existing voice codecs for 3GPP.
- ETSI DSR Extended Front-End standards ES 202 211 (XFE) and ES 202 212 (XAFE).

Recently, the SA4 codecs group within 3GPP has commissioned a comparison of the two codecs for use over a packet data channel. The study, carried out by IBM and SpeechWorks (now Scansoft) revealed a performance advantage of the DSR scheme [1].

AMR, on the other hand, was shown to exhibit very good recognition accuracy for a wide range of carrier-to-interference ratios (C/I) for transmission over a GSM circuit-switched voice channel [2]. This is probably due to the close interaction of source and channel coding, which is currently not considered in the DSR standard. The source bit rate of AMR is chosen in the range of 4.75 up to 12.2 kb/s according to the quality of the transmission channel, and the source bits are categorized in classes with different error protection capability assigned to the classes according to their relevance for decoding. The received bitstream is decoded into the speech signal at the server side, followed by feature extraction and speech recognition. This approach is known as Network Speech Recognition (NSR).

Similar ideas about joint source-channel coding have also been proposed for DSR. Weerackody et al. [3] proposed unequal error protection and a soft-feature error concealment strategy. Bernard and Alwan devised a coding scheme which allows error detection capabilities with soft-decision decoding, and the Viterbi decoder in the recognizer was modified to deal with unreliable features [4]. Peinado et al. [5] applied the concept of softbit speech decoding introduced by Fingscheidt and Vary [6] for DSR and achieved good recognition performance for AWGN and bursty channels. In their work, they assumed the bit reliability information to be known, e.g. from a given or assumed SNR-value. In [7] we built upon the same concept, however, we employed the bit reliability information as it is computed in the channel decoder. Speech recognition experiments over a simulated GSM link revealed excellent performance down to very low C/I values.

Inspired by these promising results we further explored our concept of SFDSR and present in the following results on the usefulness of intra-frame subvector correlations, database independence of the a priori information, and on some complexity issues.

Given these new developments in distributed speech recognition it is interesting to compare SFDSR with the ETSI DSR scheme and with AMR-based Network Speech Recognition (AMR-NSR). We conducted this comparison by simulating a complete GSM link, including channel (de)coding, (de)interleaving, GMSK (de)modulation and employing realistic channel models including multipath propagation, fast fading, and cochannel interference. Note that a comparison on the basis of error patterns, as is often done [5], may be inappropriate since it is unable to simulate certain joint source-channel coding concepts.

2. SOFT-FEATURE DISTRIBUTED SPEECH RECOGNITION

2.1. Concept

In this section we briefly review our Soft-Feature DSR (SFDSR) concept as illustrated in Fig. 1. For a detailed description the reader is referred to [7].

The ETSI DSR Advanced Front-End delivers features which are encoded with a split vector quantization scheme: Two feature vector components (either c_i and c_{i+1}, $i = 1, 3, \ldots, 11$, or c_0 and $\log E$) are grouped into a feature-pair subvector, and each subvector $x_n^{SV_k}$ is quantized into a bit pattern $\mathbf{b}_n^{SV_k}$ of size $M(SV_k)$ using its own codebook. Here, n is the frame index. The superscript SV_k shall denote the k-th subvector, $k = 1, \ldots, 7$, and it is sometimes omitted in the following for notational convenience.

The transmission of the bit combination is described by an equivalent channel model with input $\mathbf{b}_n$ and output $\hat{\mathbf{b}}_n$, which may

comprise the channel model itself, the channel encoder/decoder, modulation/demodulation and equalization. For soft feature speech recognition, a channel decoder is required which outputs the detected bit sequence $\hat{\mathbf{b}}_n$, and in addition reliability information in terms of estimated bit error probabilities: $\mathbf{p}_{e_n}$. We employed the MAP-decoder after Bahl et al. [8]. Using these, the transition probabilities $P(\hat{\mathbf{b}}_n|\mathbf{b}_n)$ from the transmitted bit pattern $\mathbf{b}_n$ to the known received bit pattern $\hat{\mathbf{b}}_n$ are computed. Obviously a simulation based on a description of the channel effects by error patterns is inappropriate since the bit reliability information is missing.

The next knowledge source to be exploited is the residual redundancy in the bit stream of the source coder. Here, we used *first-order* a priori knowledge $P(\mathbf{b}_n|\mathbf{b}_{n-1})$, which captures the correlation between successive frames.

The a posteriori probabilities $P(\mathbf{b}_n|\hat{\mathbf{b}}_n, \hat{\mathbf{B}}_{n-1})$ are computed from the transition probabilities and the a priori probabilities, and finally they are used to compute various parameter estimates. Here, $\hat{\mathbf{B}}_{n-1}$ is a short-hand notation for the history $\hat{\mathbf{b}}_{n-1}, \hat{\mathbf{b}}_{n-2}, \ldots$. While for speech reconstruction one is only interested in the MMSE estimates $\mu_{\hat{x}_n}$ [6], i.e. the first-order moment of the a posteriori probability mass function, the second-order moment $\sigma^2_{\hat{x}_n}$ carries additional important information which we employ for uncertainty decoding [10]. It is a measure of the confidence about the reconstructed features.

The proposed SFDSR has the following useful properties:

- It is compatible with the ETSI DSR standards: the modifications pertain only to the processing at the server side.
- The soft-feature computation is done on a subvector level. Error mitigation on a subvector-basis, as proposed in [12], is an inherent property of our approach.
- The C/I or SNR values of the channel need not be known and constant as in [5]. The channel decoder delivers the time-variant bit reliability information as a side effect of the decoding.
- Uncertainty decoding does not employ any tuneable parameter as in [4], the time-variant uncertainty information is directly computed from the a posteriori probabilities.
- The decoding does not introduce a latency as the ETSI error mitigation scheme does in the case of successive frame-pairs being affected by an error burst.

On the other hand, potential drawbacks of the system are:

- Correlation among the subvectors are ignored.
- The a prori probabilities have to be estimated in advance on a training database. This might introduce an unwanted database dependency of the achievable performance.
- The system is more computationaly demanding. In particular, a channel decoder has to be used that provides "soft-outputs" and the estimation of $\mu_{\hat{x}_n}$ and $\sigma^2_{\hat{x}_n}$ has a high computational load.

These issues are addressed in the next subsections.

2.2. Intra-Frame Subvector Correlation

Computing the a priori knowledge on a subvector level, as is done in [7], is much less complex than considering the feature vector as a whole, but potentially suboptimal since correlations among

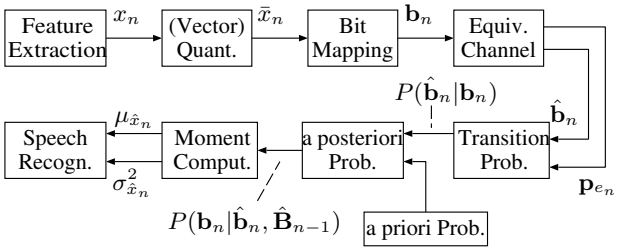

Fig. 1. *Block diagram of Soft-Feature Distributed Speech Recognition system.*

subvectors inside a frame are ignored. To check the validity of this approximation, we estimated the average mutual information $H(\mathbf{b}_n^{SV_k}) - H(\mathbf{b}_n^{SV_k}|\mathbf{b}_n^{SV_l})$ between subvectors SV_k and SV_l. Table 1 shows the mutual information for the parameters of the ETSI DSR Advanced Front-End, measured on the clean training data set of Aurora 2 [11]. The correlation between any two subvectors of cepstral coefficients is small, due to the properties of cosine transform, whereas the $(c_0, logE)$ exhibits some correlation with other subvectors, as energy is a global measure over all cepstral coefficients. Since taking into account this mutual information would increase very much the memory requirements and computational burden, we have decided not to use it.

2.3. Portability of a priori Information

We also checked whether the a priori information depends on the audio-database on which it was estimated. We computed the a priori probabilities on the Wallstreet Journal database and we used them in a SFDSR simulation involving an Aurora 2 speech recognition task. Virtually no effect on the final word accuracy of the recognizer was noticed, compared with the case when the a priori information was estimated on Aurora 2. This demonstrates that at least for a given language, english in our case, the table of a priori probabilities can be computed once and for all and need not be adapted to new applications.

2.4. Complexity Reduction

In the case of first-order a priori knowledge the probability mass functions $P(\mathbf{b}_n^{SV_k}|\mathbf{b}_{n-1}^{SV_k})$, $k = 1, \ldots, 7$ are employed to compute the MMSE estimates of the parameters, and, in case of uncertainty decoding, also to compute the variance to be used in the recognizer.

The table $P(\mathbf{b}_n^{SV_k}|\mathbf{b}_{n-1}^{SV_k})$ consists of $2^{M(SV_k)} \times 2^{M(SV_k)}$ entries. This amounts to a total of $5 \cdot 2^{12} + 2^{10} + 2^{16} = 87,040$ values for the seven subvectors SV_1 to SV_7. A closer look at the table, however, reveals that about 50% of the entries are zero. It is therefore advantageous to store only the nonzero values in a linked list, avoiding unnecessary multiplications with zero in the MMSE parameter estimation. Due to the quadratic dependence of the number of multiplication on the number of nonzero terms, this simple modification reduces the computational effort of the parameter estimation by a factor of four.

Note that a zero value of $P(\mathbf{b}_n^{SV_k}|\mathbf{b}_{n-1}^{SV_k})$ indicates that the bit pattern $\mathbf{b}_n^{SV_k}$ can never follow $\mathbf{b}_{n-1}^{SV_k}$. This information could be used as a more elaborate consistency check compared to the one

Table 1. *Entropies $H(\mathbf{b}_n^{SV_k})$ and mutual informations $I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_l}) = H(\mathbf{b}_n^{SV_k}) - H(\mathbf{b}_n^{SV_k}|\mathbf{b}_n^{SV_l})$ among the subvectors of the ETSI advanced DSR front-end.*

Subvector SV_k	$SV_1 = c_1, c_2$	$SV_2 = c_3, c_4$	$SV_3 = c_5, c_6$	$SV_4 = c_7, c_8$	$SV_5 = c_9, c_{10}$	$SV_6 = c_{11}, c_{12}$	$SV_7 = c_0, \log E$
$M(SV_k)$	6	6	6	6	6	5	8
$H(\mathbf{b}_n^{SV_k})$	5.84	5.80	5.80	5.77	5.82	4.80	7.72
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_1})$	-	0.32	0.17	0.16	0.11	0.06	1.28
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_2})$	0.32	-	0.12	0.08	0.06	0.07	0.40
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_3})$	0.17	0.12	-	0.04	0.04	0.05	0.23
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_4})$	0.16	0.08	0.04	-	0.07	0.04	0.18
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_5})$	0.11	0.06	0.04	0.07	-	0.06	0.14
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_6})$	0.06	0.07	0.05	0.04	0.06	-	0.11
$I(\mathbf{b}_n^{SV_k};\mathbf{b}_n^{SV_7})$	1.28	0.40	0.23	0.18	0.14	0.11	-

described in the ETSI DSR standard. Using the thresholds defined in [14] to prune those entries which do not pass the consistency test, reduced the number of useful entries to 20% , however the recognition accuracy was significantly decreased.

Concerning the use of the complex Bahl channel decoder, it can be stated that a "soft-output" decoder is required anyway if turbo codes have to be used as is the case in UMTS data transmission. Further, the less complex soft-output Viterbi algorithm (SOVA) can do as well [9].

3. COMPARISON OF CODEC OPTIONS

3.1. Experimental Setup

In the following we present experimental results on a comparison of three approaches: AMR-NSR, ETSI DSR and Soft-Feature DSR:

- Network-based speech recognition employing the AMR speech codec at source data rates of 4.75, 5.9, 7.4 and 10.2 kb/s. The corresponding channel coding was applied as described in the specifications [15] and the signal was transmitted over a voice channel. At the server side the speech signal was reconstructed, features according to the advanced front-end feature extraction algorithm were computed and fed into the recognizer. This approach is called "AMRx" in the following, where "x" denotes the chosen source bit rate.
- DSR according to XAFE. The 5.6 kb/s source bitstream was transmitted via a GSM data channel (TCH/F4.8 [15]) which uses convolutional coding with a rate $r = 1/3$.
- Soft-Feature DSR as described in [7] (SFDSR). The 4.8 kb/s source bit stream was transmitted over the same channel as in the DSR-XAFE case. At the server side, MMSE estimates of the features and uncertainty information were computed and both fed to the recognizer for uncertainty decoding.

In all three scenarios the gross bit rate was 22.8 kb/s, corresponding to a GSM full rate traffic channel.

We employed the GSM library of the SPW ("Signal Processing Worksystem") software suite to simulate the physical layer of the GSM link. The simulation of the GSM transmission consisted of the following elements

- Interleaving at the transmitter and deinterleaving at the receiver side
- Channel model approximating the COST 207 profile: a "typical urban" channel, modelled by 12 propagation paths (delay spread 1.03 μs) and Rayleigh fading. The mobile terminal was assumed to be moving at 50 km/h. Further, cochannel interference was simulated at various C/I (carrier-to-interference) ratios.
- Channel decoding with the Bahl algorithm which delivers bit reliability information required for MMSE parameter estimation.

In our simulations, however, we assumed perfect synchronization of GSM system components and no link adaptation was involved in the case of AMR.

3.2. Experimental Results

In the following we present results about speech recognition experiments on the clean data set of AURORA 2 database.

Figure 2 shows the achieved word accuracy as function of C/I ratio. SFDSR can be seen to clearly outperform the other approaches at low C/I values. As was expected, in the case of a noisy channel, the lower the source coding rate the better the recognition performance of the AMR-based systems. This is due to the more powerful channel codes that can be accommodated in the data stream of fixed gross bit rate of 22.8kb/s.

Note that for low C/I values, AMR475 and AMR59 even outperform DSR. This fact cannot be seen in figure 3 where the word accuracy is shown as a function of bit error rate (BER). Only the presentation as a function of C/I, as is chosen in figure 2 reveals the effect of the different channel coding schemes.

Comparing the two figures it can therefore be concluded that DSR is superior to AMR given the same bit error rate. However, the specified channel codes for the low bit rate AMR modes are more powerful, resulting in a performance advantage over DSR at low C/I values.

4. CONCLUSIONS

In this paper we compared Soft-Feature DSR with the two candidate codecs for Speech Enabled Services and showed that it was

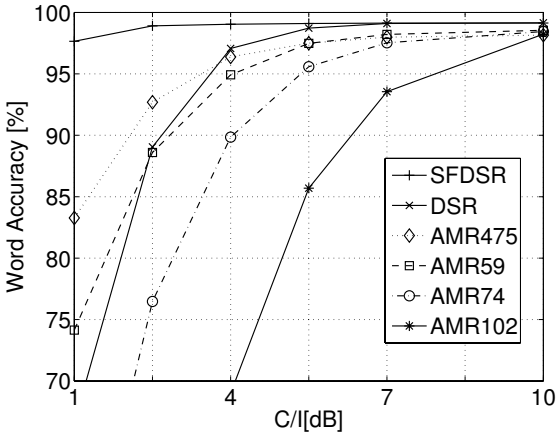

Fig. 2. *Word accuracy vs C/I ratio.*

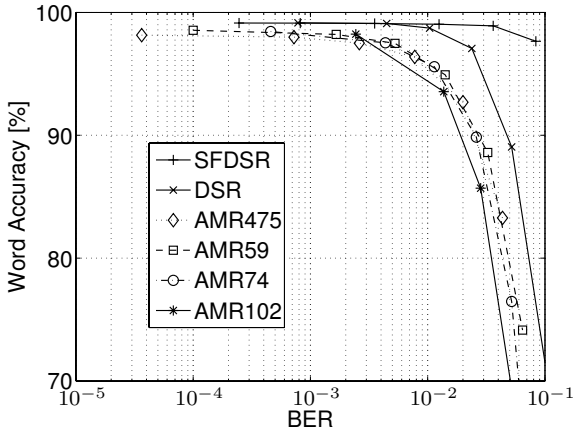

Fig. 3. *Word accuracy vs bit error rate (BER).*

superior to both of them. The comparison was made in terms of achieved recognition word accuracy, assuming the transmission of speech parameters over an error prone circuit-switched communication network.

For the same bit error rate it was shown that the ETSI DSR scheme outperforms the AMR-based network speech recognition. However, simulations of the GSM physical channel also revealed that the specified channel codecs to be used with AMR may result in improved recognition performance at low C/I, compared to the ETSI DSR method.

5. ACKNOWLEDGEMENTS

This work was supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/2-1.

6. REFERENCES

[1] D. Pearce, "Enabling speech & multimodal services on mobile devices: the ETSI Aurora DSR standards & 3GPP Speech Enabled Services", in *Proc. AVIOS/SpeechTEK Spring 2004 Conference*, San Francisco, May 2004.

[2] T. Fingscheidt et al., "Network-based vs. distributed speech recognition in adaptive multi-rate wireless systems", in *Proc. ICSLP 2002*, Denver, Co., Sep. 2002.

[3] V. Weerackody, W. Reichl, and A. Potamianos, "An error protected speech recognition system for wireless communications", *IEEE Trans. on Wireless Communications*, vol. 1, pp. 282-291, 2002.

[4] A. Bernard and A. Alwan, "A Low-bitrate distributed speech recognition for packet-based and wireless communications", *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 570-579, 2002.

[5] A. Peinado et al., "HMM-based channel error mitigation and its application to distributed speech recognition", *Speech Communication*, vol. 41, pp. 549-561, 2003.

[6] T. Fingscheidt, P. Vary, "Softbit speech decoding: A new approach to error concealment", *IEEE Trans. on Speech an Audio Processing*, vol. 9, no. 3, pp. 1-11, 2001.

[7] R. Haeb-Umbach, V. Ion, "Soft features for improved distributed speech recognition over wireless networks", in *Proc. ICSLP 2004*, Jeju, Korea, Oct. 2004.

[8] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate", *IEEE Trans. on Information Theory*, vol. 20, pp. 284-287, March 1974.

[9] J. Hagenauer, P. Höher, "A viterbi algorithm with soft-decision outputs and its applications", in *Proc. IEEE Global Communications Conference*, Dallas, Tx, Nov. 1989.

[10] L. Deng, J. Droppo, and A. Acero, "Exploiting variances in robust feature extraction based on a parametric model of speech distortion", in *Proc. ICSLP 2002*, Denver, Co., Sep. 2002.

[11] H.G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *ISCA ITRW Workshop ASR2000*, Paris, France, Sep. 2000.

[12] Z. Tan, P. Dalsgaard, and Borge Lindberg, "A subvector-based error concealment algorithm for speech recognition over mobile networks", in *Proc. ICASSP 2004*, Montreal, CA., May 2004.

[13] ETSI EN 301 704 V7.2.1 "Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1 Release 1998)" Apr. 2000

[14] ETSI ES 202 212 v1.1.1 "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm.", Nov. 2003

[15] ETSI TS 100 909 v8.7.1 "Digital cellular telecommunications system (Phase 2+); Channel coding (3GPP TS 05.03 version 8.7.0 Release 1999)", Apr. 2003