Harmonic Distortion in the TETRA Channel and its Impact on Automatic Speech Recognition

Daniel Stein, Thomas Winkler, and Jochen Schwenninger
Fraunhofer Institute for Intelligent Analysis and Information Systems
Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Introduction

Since its introduction in the mid 90ies, the Terrestrial Trunked Radio (TETRA) standard for digital radio broadcast has been deployed in most European and Asian government networks and public safety networks. The underlying Adaptive Code-Excitation Linear Prediction scheme employs adaptive code books which are optimized for human speech. However, harmonic distortions of the hardware are amplified as well, which we will show via spectrograms and frequency analysis. While studies indicate that human perception neglects these distortions up to a Total Harmonic Distortion of about 1% and sometimes even above (e.g. [1]), their impact on Automatic Speech Recognition (ASR) can be dramatic. In this paper, we demonstrate this effect on a strong recognition system for German broadcast news, using a dedicated fire fighter radio transmitter. We will dissect the influence of the hardware and the software components by analysing the word error rate, typical word substitutions and changes in the extracted features.

TETRA Codec

Terrestrial trunked radio (TETRA) [2] is a standard for a digital trunked radio system, first published by the European Telecommunications Standards Institute (ETSI) in 1995. It has been designed for robust speech transmission and indeed is used, e.g., for public safety networks across Europe, Asia and other countries. However, its influence on ASR has rarely been analysed.

Scientific papers analysing the impact on natural language processing by automatic means are scarce. [3] analyse the TETRA codec on the speaker recognition performance. They do not only work on the audio signal, but also make direct use of the linear prediction coefficients that are computed by the TETRA encoder. Simply taking the decoded speech signal performs worst and seems to be the hardest setting. [4] is one of the few papers employing actual TETRA data in their recognition setup. On a small corpus of spoken German digits, they show that the TETRA codec performs poorly in comparison to the plain signal, to a 16 kbit/s Code-Excited Linear Prediction (CELP), and to a GSM codec.

Preliminaries

We use the ASR system as described in [5], which is a state-of-the-art recognizer for German and English. For training and testing we employ two manually transcribed, distinct sets of German broadcast news and political talkshows. The original audio is sampled at 16 kHz and can be considered to be of clean quality. Noisy sections of the recordings have been omitted. The training set consists of 82 799 sentences (723 933 running words, 52 100 distinct), and the test set consists of 5 719 sentences (46 978 running words and 8 799 distinct words).

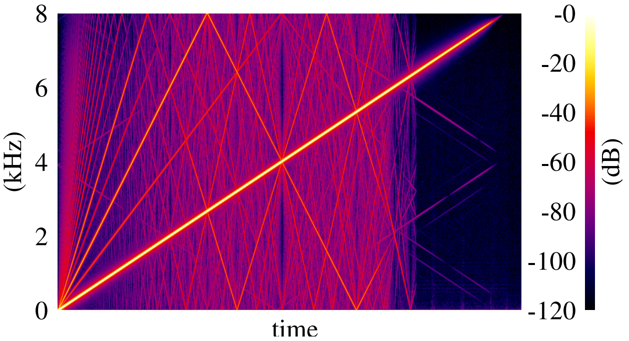

Figure 1: Spectrogram of the test signal in a closed-loop setting. Common harmonic distortions appear.

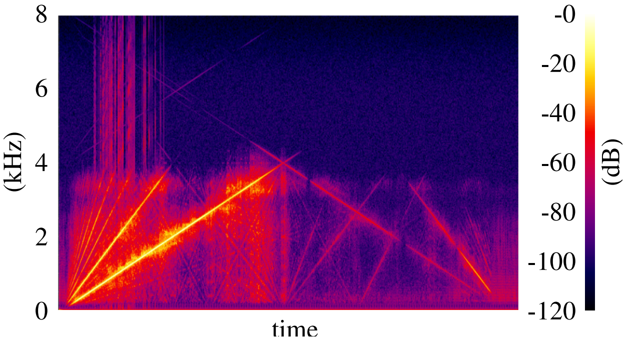

Figure 2: Spectrogram of the test signal as received over TETRA

We employ the CM 5000 radio station and the MTP 850 handheld device, both by the Motorola$^{\text{TM}}$ Corporation. To characterize the frequency characteristics of the actual TETRA channel, we transmitted a synthesized frequency sweep from 0 to 8 kHz. In the setup, the CM 5000 is used as sender, having the input signal fed in via the headset connector. The MTP 850 acts as the receiver, and the signal is recorded from the line-out. Figure 2 illustrates the drastic quality deterioration: the encoding and subsequent transmission adds noise to the whole spectrum and suppresses all frequencies above 4 kHz. In a separate experiment where we re-recorded the signal without TETRA transmission, we could attribute the massive amounts of harmonic distortion as witnessed in the spectrogram to the audio hardware.

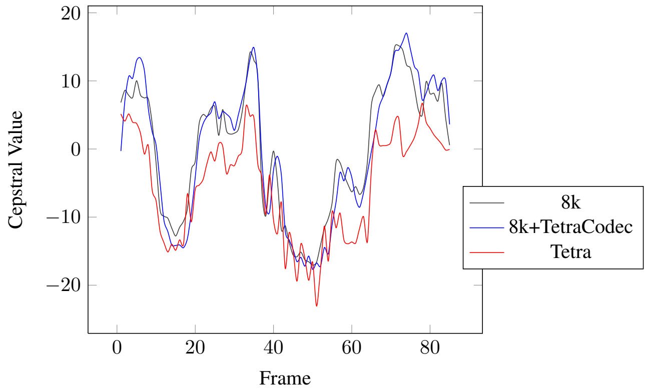

Figure 3: Developing of 2^{nd} cepstral coefficient of the phrase "Danke Karen"

Separation of Recognition Influences

We conducted a set of experiments where we process both training and test set in the same manner, to separate the effects that lead to the recognizer performance drop. The results can be seen in Table 1. The word error rate (WER) for the clean speech (16 kHz) is at 26.6% and at 42.3% for the TETRA-clean signal. Based on this data set, we attribute 2.5% to the frequency low-pass effect from 16 to 8 kHz. Another 6.5% absolute can probably be explained by the ACELP procedure within the TETRA encoding scheme. This can be witnessed when applying the conceptually similar AMR 4.75 codec with the same bandwidth as TETRA to the resampled data. The additional processing inside the TETRA codec itself does not seem to degrade the performance substantially and only adds another 1.8% absolute WER. From this TETRA codec result, the actual influence of the broadcast station can be measured at an additional 4.9% WER degradation.

In Figure 3 we can examplarily see the 2^{nd} coefficient for the 8 kHz clean speech, the TETRA codec and the TETRA radio station signal. While the coefficients for clean speech and TETRA codec are rather similar, the same coefficient for the TETRA radio signal shows two major differences. As we have some additional tenth of seconds before and after the speech signal to avoid cutting the speech while recording, feature extraction and cepstral normalisation is performed on a larger set of cepstral feature vectors. Thus, the features are not necessarily normalised for the aligned frames only in Figure 3. Furthermore, some additional distortion of the TETRA radio coefficient is obvious, especially between frame 40 and 60, which we assume is caused by the harmonic distortion.Both effects probably explain the 20.5% absolute drop in WER from TETRA radio data tested on the TETRA codec models compared to the same data tested on the TETRA radio models.

The main substitution errors of TETRA radio test on TETRA codec models are based on phoneme confusion of "i" and "e" as well as "m" and "n" probably caused by the harmonic distortion. Other errors include a deletion of phonemes like "s" and "d". However, these are presumably erroneous because of the low-pass effect of TETRA.

Table 1: Performance loss through TETRA codecs. "Clean" is the original signal. "AMR 4.75 and "TETRA codec" refer to the respective software codec applied to a clean signal. "TETRA radio" refers to the signal looped through TETRA hardware.

condition train	condition test	WER in %
clean 16 / 8 kHz	clean 16 / 8 kHz	26.6 / 29.1
AMR 4.75	AMR 4.75	35.6
TETRA codec	TETRA codec	37.4
TETRA radio	TETRA radio	42.3
TETRA codec	TETRA radio	62.8

Conclusion

In this paper, we offered a detailed analysis of the TETRA channel impact on automatic speech recognition performance. We highlighted which aspect of the channel can be contributed the most to the performance drop. It seems that, even though the TETRA codec has been designed with speech intelligibility in mind, it emphasizes harmonic distortions introduced by the radio equipment, which severly hampers common automatic speech recognition systems. In a next step, we are going to address this problem by investigating how to counteract this phenomenon by, e.g., adding artificial distortion to the training material.

This work has been partly funded by the European Community's Seventh Framework Programme (FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics – STREP) under grant agreement n° 231738.

References

[1] Earl R. Geddes and Lidia W. Lee, "Auditory perception of nonlinear distortion - theory," in *Audio Engineering Society Convention 115*, Oct. 2003.

[2] ETSI, "Terrestrial trunked radio (tetra); speech codec for full-rate traffic channel; part 2: Tetra codec," Tech. Rep. ETS 300 395-2, European Telecommunication Standard, Feb. 1998.

[3] Alexandre Preti, Bertrand Ravera, François Capman, and Jean-François Bonastre, "An Application Constrained Front End for Speaker Verification," in *Proc. of the 16th European Signal Processing (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[4] S. Euler and J. Zinke, "The Influence of Speech Coding Algorithms on Automatic Speech Recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1994, vol. i, pp. I/621–I/624.

[5] D. Schneider, J. Schon, and S. Eickeler, "Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System," in *Proc. SIGIR*, Singapore, 2008.