TETRA CHANNEL SIMULATION FOR AUTOMATIC SPEECH RECOGNITION

Daniel Stein, Thomas Winkler, Jochen Schwenninger, and Rolf Bardeli

Fraunhofer Institute for Intelligent Analysis and Information Systems
Schloss Birlinghoven, 53754 Sankt Augustin, Germany

ABSTRACT

The problem of deterioated automatic speech recognition performance for telecommunication devices like mobile phones is well-studied. However, few papers analyse the influence of emergency broadcast radio devices such as the Terrestrial Trunked Radio (TETRA), which is used for many public safety networks in Europe and Asia. In this paper, we dissect several aspects of the TETRA encoding scheme when transmitting over a dedicated emergency radio station by conducting experiments with original hardware and corresponding software simulations. We also offer a new technique for a cepstral mismatch compensation.

1. INTRODUCTION AND RELATED WORK

Terrestrial trunked radio (TETRA) [1] is a standard for digital trunked radio systems, first published by the European Telecommunications Standards Institute (ETSI) in 1995. It has been designed for robust speech transmission and indeed is used for public safety networks across Europe, Asia and other countries. However, its influence on automatic speech recognition (ASR) has rarely been analysed.

The objective speech quality of the TETRA codec is examined in [2]. While focussing more on the speech quality degradation in correlation with the bit error rate, the findings are, as the authors mention themselves, somewhat inconclusive. [3] offers an extensive overview of the TETRA encoding/decoding performance, and on package delay and throughput in an overall architecture, with special focus on transmission errors and co-channel interference.

Scientific papers analysing the TETRA encoding impact on natural language processing by automatic means are scarce. [4] analyses 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. [5] 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.

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.

Fig. 1. Motorola's CM 5000 radio station, as used in the experiments

In this paper, we report on several experiments to identify the challenges encountered by an ASR backend that only has access to TETRA output. For this, we employ a TETRA broadcast station as used by the German firefighters and analyse the impact of the setup on a medium-sized German TV broadcast corpus. We further analyse the influence of training material preprocessing and additional noise, devise strategies for adapting acoustic models without access to TETRA radio output, and evaluate several frontends for feature extraction.

2. TECHNICAL BACKGROUND

In this section, we briefly review the TETRA encoding scheme, as well as harmonic distortions, which seem to play a major role in ASR on TETRA signals. For the hardware, we employ the CM 5000 radio station (Figure 1) and the MTP 850 handheld device, both by Motorola.

2.1. TETRA Encoding

The TETRA speech codec is based on the CELP coding model. It employs both a short-term synthesis filter working with linear prediction coefficients and a pitch filter working with an adaptive codebook. For a set of linear prediction coefficients a_i with order $p = 10$, the short-term synthesis filter is given by:

$$\frac{1}{S(z)} = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}} \,. \tag{1}$$

For a pitch delay T and a pitch gain g_p, the pitch filter is given by:

$$H(z) = \frac{1}{P(z)} = \frac{1}{1 - g_p z^{-T}} \,. \tag{2}$$

Pitch and excitation codebook parameters are determined by selecting the candidate that has the closest output to the perceptually weighted input signal, given by the filter:

$$W(z) = \frac{S(z)}{S(z/0.85)} \,. \tag{3}$$

For the codebooks, the Algebraic CELP technique is used, i.e., the codebook vectors of the TETRA codec are fixed, but shaped according to a dynamic matrix that depends on $S(z)$, given by the Toeplitz lower triangular matrix that is constructed from the filter impulse response:

$$F(z) = \frac{S(z/0.75)}{S(z/0.85)} \,. \tag{4}$$

For a given speech signal in 8 kHz, the linear prediction coefficients are computed for each frame of 30 ms, whereas pitch and the algebraic codebook parameters are transmitted for four sub-frames of length 7.5 ms. The final bit rate is 4.567 kbit/s. For a complete overview, see [1].

2.2. Harmonic Distortion

Harmonic distortion adds full number multiples of existing frequencies to a signal. A common measure for the level of distortion is the Total Harmonic Distortion (THD) defined as the ratio of the sum of the powers P_n with $n \geq 1$ of all N multiples of a frequency to the power of the fundamental frequency P_0:

$$\text{THD} = \sum_{n=1}^{N} P_n / P_0 \tag{5}$$

Harmonic distortion can be caused by amplifiers, microphones, loudspeakers and other devices. Even high quality amplifiers often have a THD of up to 1% in the relevant spectral range. While even a large THD can be acceptable in terms of perception [6], already smaller distortions change the frequency characteristics and as a result the cepstral coefficients used for ASR.

Preliminary tests [7] not only indicate that we have to expect harmonic distortion caused by our TETRA radio equipment, but also that the adaptive code book in the TETRA codec emphasizes them. The reason might well be that, since the codec is optimized on human speech intelligibility, special focus is on the preservation of harmonics produced in voiced phonemes, at the cost of further amplified harmonic distortions in the signal. We will analyse this further in Section 5.2.

3. CEPSTRAL MISMATCH COMPENSATION

In [8] a Lombard effect and noise compensation method is introduced. The authors show that distortion in the spectral domain caused by the Lombard effect as well as noise can be described and compensated in the cepstral domain as shown in the following equation:

$$c_{\text{clean},n} = \sum_{k=0}^{K} A(n,k) c_{\text{Lombard},k} + b(n) \tag{6}$$

The n^{th} cepstral value $c_{\text{clean},n}$ of clean speech can be estimated from all K cepstral values $c_{\text{Lombard},k}$ of a comparable frame of Lombard speech. The coefficients $A(n,k)$ and $B(n)$ are determined using multiple linear regression (MLR) based on all comparable frame pairs of clean and Lombard speech.

We assume that the described compensation can also be used to formulate an approximation of a general transformation from the cepstral vector $\mathbf{c}_{\text{utt}}$ of the test utterance to the "clean" cepstral vector $\mathbf{c}_{\text{am}}$ of the acoustic model. To enable real-time applications we reduce the full set of our acoustic models with about 200,000 mean vectors to a codebook of 300 cepstral mean vectors with k-means. We expect that similar acoustic units and mixture components of the same unit form clusters in the feature space, so that the quantisation error using the codebook is rather low.

Now, we determine the presumably corresponding vector i of the codebook $\tilde{\mathbf{c}}_{\text{am},i}$ to the m^{th} frame's feature vector $\mathbf{c}_{\text{utt},m}$ of the utterance by minimum Euclidean distance. Further, we only consider vector pairs with a distance below the mean distance of all pairs of an utterance ("reliable pairs") to avoid compensation with possibly incorrectly assigned pairs.

Equivalent to Equation 6 we can now estimate the tranformation coefficients $A(n,k)$ and $b(n)$ for all reliable pairs. Assuming that we have a sufficient number of reliable pairs, structural differences should be covered by the transformation while more random differences (e.g. caused by the quantisation error) should mainly increase the overall minimum error during error minimisation of MLR.

With coefficient matrix $\mathbf{A}$ and coefficient vector $\mathbf{b}$, we can now estimate the compensated cepstral feature vectors $\hat{\mathbf{c}}_{\text{utt},m}$ of all frames m of the utterance by using the following equation similar to CMLLR feature adaptation:

$$\hat{\mathbf{c}}_{\text{utt},m} = \mathbf{A}\mathbf{c}_{\text{utt},m} + \mathbf{b} \tag{7}$$

For ASR the compensated feature vectors $\hat{\mathbf{c}}_{\text{utt},m}$ are used.

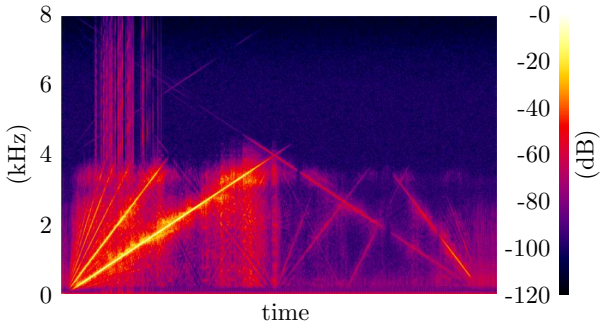


Fig. 2. Spectrogram of a sweep test signal as received over TETRA

4. PRELIMINARIES

To characterise the frequency characteristics of the actual TETRA channel, we transmitted a synthesised 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 real TETRA transmission, we could attribute the massive amounts of harmonic distortion as witnessed in the spectrogram to the audio hardware.

Since we felt that such a clean signal is not to be expected in real-life settings, especially in rescue operations, we also created a second setup of experiments. It introduces distortion caused by non-ideal microphones and ambient noises. To achieve this, we feed the signal into the sender via the MTP 850's internal microphone placed in front of quality loudspeakers. This setup resembles real-life usage of the TETRA channel very closely.

In this paper, we denote the acoustic signal that is directly recorded from line-out as *TETRA-clean*, and the signal further distorted by the microphone and the loudspeaker as *TETRA-noise*.

For feature extraction, we employ the HTK toolkit,[1] and extract 39 features (12 MFCCs with energy, plus deltas and accelerations, using zero-mean) for each frame of 25 ms window length using a stepsize of 10 ms. As an alternative, we extract 12 features using the ETSI advanced frontend,[2] but compute the deltas and accelerations with HTK using the previous configuration. We use the TETRA codec reference implementation as provided by ETSI.[3] The Adaptive Multi-Rate (AMR)[4] speech codec has been used in the experiments since it features a comparable ACELP scheme. Recording and resampling of the audio signals was carried out using SoX.[5] For language modeling, we make use of the MIT Language Modelling Toolkit[6] to compute a trigram language model with modified Kneser-Ney smoothing. We use the Julius toolkit[7] for decoding.

For training and testing we employ two manually transcribed, distinct sets of German broadcast news and political talk-shows. 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).

[1] htk.eng.cam.ac.uk/

[2] www.etsi.org/WebSite/Technologies/DistributedSpeechRecognition.aspx

[3] pda.etsi.org

[4] www.3gpp.org/ftp/Specs/html-info/26104.htm

[5] sox.sourceforge.net

[6] code.google.com/p/mitlm/

[7] julius.sourceforge.jp/

5. EXPERIMENTS

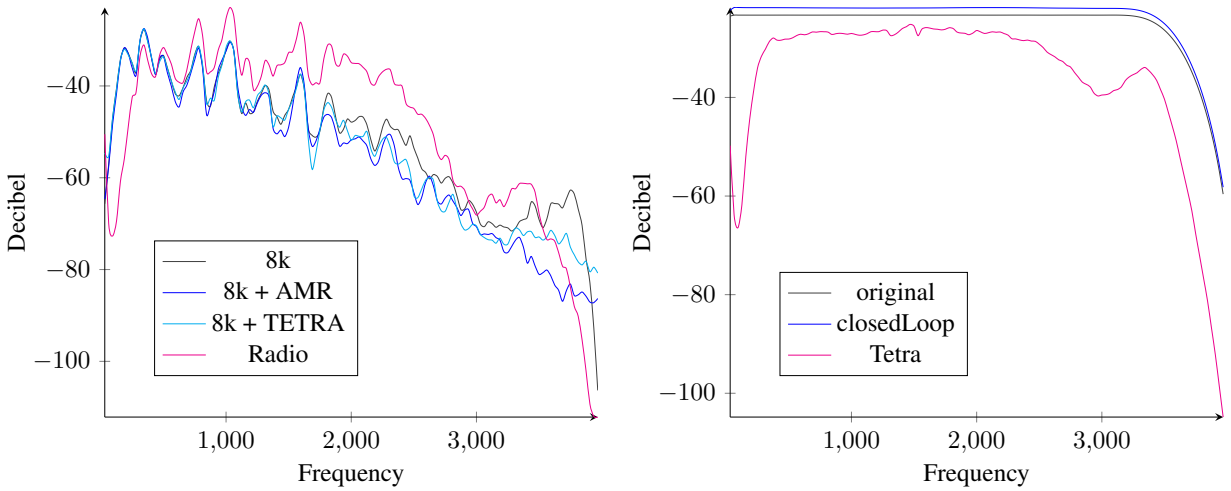
We performed three sets of experiments which analyse various aspects of the TETRA transmission channel and their impact on ASR performance, such as signal pre-processing, suitable feature extraction and additional noise.

5.1. 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 recogniser performance drop. The results are given in Table 1. The word error rate (WER) for the clean speech is at 26.6 WER and at 42.3 WER for the TETRA-clean signal. Based on this data set, we attribute 2.5 WER to the frequency low-pass effect. Another 6.5 WER 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 WER. From this TETRA codec result, the actual influence of the broadcast station can be measured at an additional 4.9 WER degradation. The effect of the different preprocessing steps in comparison to the real radio signal is visualised in the spectrogram in Figure 3(a).

5.2. Simulating a real TETRA radio channel

In this set of experiments, we always measure the ASR performance on the TETRA-clean test set. Further, we assume that only the test data is available from a real TETRA device, and the acoustic models should be fitted as well as possible using channel simulations. The ideal case is to transmit all training data through a real TETRA radio channel, as performed above, which leads to a WER of 42.3. Table 2 shows the ASR



(a) Frequency analysis for the word "Tagesthemen"

(b) Frequency analysis on synthetic sweep 0-4 kHz

Fig. 3. Influence of the channels on the frequencies

Table 1. Performance loss through TETRA codecs. Same preprocessing of training & test material. HTK was used as frontend.

kHz	codec (train & test)	WER
16	–	26.6
8	–	29.1
8	AMR 4.75	35.6
8	TETRA (codec)	37.4
8	TETRA-clean (hardware)	42.3

performance for a number of approximations. A huge gain can of course be seen by matching the sample rate: resampling the training data to 8 kHz results in a 13.9% decrease of the WER.

What is more interesting is that by applying the codecs from Section 5.1 afterwards, this only results in minor performance boosts of 0.9% for AMR 4.75 and 0.6% absolute for TETRA. Further analysing which effect causes this huge discrepancy between the best WER of 62.5 — still 20.2 points absolute worse than using acoustic models trained on real TETRA data — we investigated the influence of artificially adding channel noise, equalization effects and total harmonic distortion (see Table 3), this time by relying on the hardware equipment.

First, we checked the channel noise as introduced by the equipment, clearly audible as a buzzing sound when silence is transmitted. We recorded this "silence noise" and mixed it into the training material. This led only to a marginal reduction in WER of 0.3 absolute. Next, we checked the influence of non-linear frequency responses in real transmisson setups. Measuring the frequency response by recording a synthetic sweep from 0 to 4 kHz as shown in Figure 3(b), parameters for simulating the equalization were obtained and applied to the training data. Resulting in a WER of 64.1, the effect is apparently negligible.

It seems that the major contribution to the error rate is due to the harmonic distortion effects as witnessed in Figure 2. We simulated this effext by adding harmonic distortions to the clean signal, as follows: for each bin b in the spectrogram, we have shifted its frequency content to the bins nb where $n \in \{2, 3, \ldots, 21\}$, i.e., we have added 20 harmonics to the original signal. The shifting was performed in the frequency domain (128 ms window length, 7/8-th overlap) and phase has been corrected such that phase angle ϕ in bin b has been transformed to phase angle $n\phi$ in bin nb. The strength of the individual harmonics has been calibrated by the intensity measured in a frequency sweep that has been transmitted via TETRA radio. This has led to much stronger harmonics than those found in speech signals transmitted via TETRA. Therefore, we have attenuated each harmonic by a factor of 0.02 which led to similar harmonic distortions as in real TETRA transmissions.

As can be seen in Table 3, the added harmonics improve to 60.1 WER, an increase of 3.3 absolute. From the distortions simulated, this is the largest increase. This strongly suggests that THD introduced by the equipment and probably further emphasised by the codec, contribute most to the large performance degradation. While for a human ear this effect might be inaudible, it heavily affects the MFCC balance. Manually checking the substitution errors, most of them are indeed based on overtone confusion (e.g. the German "u" and the German "ü").

5.3. Robustness/Frontend

For the third set of experiments, the performance of the ASR system was evaluated on the more realistic TETRA-noise setup as described in Section 4. Two additional frontends for feature extraction in noisy environments are included. As

Table 2. ASR results using channel simulation, for the TETRA-clean setting, where the test set is recorded from the TETRA radio station.

kHz	codec on train	WER
16	–	77.3
8	–	63.4
8	AMR 4.75	62.5
8	TETRA	62.8
8	TETRA clean	42.3

Table 3. ASR results on the TETRA-clean set, adding possible effects of the TETRA hardware to the clean speech.

kHz	distortion of train	WER
8	–	63.4
8	tetra equipment channel noise	63.1
8	equalization	64.1
8	artificial harmonic distortion	60.1

can be seen from Table 4, the task is even more challenging, with a baseline performance of 81.9 WER. Independent from the channel used, usage of the ETSI advanced frontend consistently degrades performance, in the case of AMR 4.75 more than 8% absolute. We assume that this is caused by the noise suppression in the frontend, which probably cancels out relevant parts of the signal and thus results in loss of valuable information. Cepstral mismatch compensation (CMC) in contrast improves performance in all settings and is the best alternative when no TETRA radio material is available. Unsurprisingly, the best results are achieved when training the acoustic models on TETRA radio material.

6. 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 contributes most to the performance drop. It is interesting to note that, while the broadcast radio station (i.e. TETRA-clean) only adds a small error on a TETRA encoded signal, it is very hard to simulate the distortion of the station without access to the radio hardware. On a realistic setting including additional noise, the recognition rate is very poor. With cepstral mismatch compensation, however, we were able to introduce a feature compensation method that extends the HTK extraction and surpasses the ETSI frontend in this particular setting.

As we had only access to one specific TETRA station and handheld device, evaluations with other TETRA equipment are necessary to generalise the results in this work.

Table 4. ASR results with different frontends, for the TETRA-noise setting, where the test set is recorded via microphone and then passed through the TETRA radio. Sampling rate is 8 kHz for all settings, results are given in WER.

codec on train	HTK	ETSI	CMC
–	81.9	85.0	75.1
AMR 4.75	74.8	83.4	74.7
TETRA	76.7	84.4	74.0

7. REFERENCES

[1] 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.

[2] C.H. Slump, T.IJ.A. Simons, and K.A.Verweij, "On the Objective Speech Quality of TETRA," in *Proc. of the Anual workshop on Circuits, Systems and Signal Processing*, Mierlo, the Netherlands, Nov. 1999, pp. 421–429.

[3] Martin Steppler, *Leistungsbewertung von TETRA-Mobilfunksystemen durch Analyse und Emulation ihrer Protokolle*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, July 2002.

[4] 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.

[5] 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.

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

[7] D. Stein, T. Winkler, and J. Schwenninger, "Harmonic Distortion in the TETRA Channel and its Impact on Automatic Speech Recognition," in *Proc. DAGA 2012*, Darmstadt, Germany, Mar. 2012, pages accepted.

[8] Sang-Mun Chi and Yung-Hwan Oh, "Lombard effect compensation and noise suppression for noisy lombard speech recognition," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4, pp. 2013–2016 vol.4, Oct 1996.