Automatic Speech Recognition on Firefighter TETRA broadcast

Daniel Stein and Bela Usabaev

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

Abstract

For a reliable keyword extraction on firefighter radio communication, a strong automatic speech recognition system is needed. However, real-life data poses several challenges like a distorted voice signal, background noise and several different speakers. Moreover, the domain is out-of-scope for common language models, and the available data is scarce. In this paper, we introduce the PRONTO corpus, which consists of German firefighter exercise transcriptions. We show that by standard adaption techniques the recognition rate already rises from virtually zero to up to 51.7% and can be further improved by domain-specific rules to 47.9%. Extending the acoustic material by semi-automatic transcription and crawled in-domain written material, we arrive at a WER of 45.2%.

Keywords: TETRA, ASR, firefighter recordings, PRONTO

1. Introduction

Keyword extraction on firefighter radio communication can be of valuable assistance, e.g. by automatically displaying material on a map interface or by assisting in the required man-power estimation. In scenarios of large-scale emergencies, the radio operator is in charge of transcribing various information submitted over the public safety network and passing the written information to the operation control. Even nowadays, this is mostly done manually. While for safety reasons a human should always be in charge of this transcription procedure, reliable keyword extraction taken from a strong automatic speech recognition system can greatly speed-up and enhance this step.

In this paper, we introduce the PRONTO firefighter database, which features real-life transcriptions from fire protection exercises. The corpus poses several challenges: audio-wise, we have to cope with heavy channel distortion from the radio station, background noise, different speakers, and a local dialect. Domain-wise, the issues include an open vocabulary, different grammar, technical terms and, because a two-way radio system is employed, heavy use of voice procedure (e.g. "affirmative", "over and out"). Moreover, the available data is relatively scarce.

We show that adaptation of the acoustic models, language model adaptation and low-pass filtering already leads to promising results. We further employ dialectal pronunciation, simple rule-based extension of the development set and post-processing steps for German compound words. In a last step, we analyse in how far new data improves the recognition performance. All methods are evaluated with the Word Error Rate (WER) on a with-held test set.

1.1. Related Work

Since its introduction in (ETSI, 1998), the Terrestrial trunked radio (TETRA) has been adopted in most European and Asian networks as the default codec within public safety networks. It is therefore also used in the exercises that we recorded. The TETRA codec emphasizes security and robustness, while maintaining a relatively low bit rate that attempts to keep human speech as intelligible as possible. While there have been many studies that investigate radio systems employing TETRA in terms of bit error rates, package loss and co-channel interference, its impact on automatic speech recognition has rarely been investigated.

(Slump et al., 1999) investigated the codec quality when it was introduced in the public safety network of the Netherlands. 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. (Steppler, 2002) offers an extensive overview of a TETRA system performance, on features like e.g. package delay and throughput, with special focus on transmission errors and co-channel interference.

However, the scientific papers on the impact on natural language processing by automatic means are scarce.

(Preti et al., 2008) 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. Among their set of experiments, the setting with the decoded speech signal performs worst and seems to be the hardest setting.

(Euler and Zinke, 1994) 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 Predictive (CELP) and to a GSM codec.

1.2. Paper Structure

This paper is organized as follows: in Section 2., we describe the TETRA codec, the software and the hardware that were employed. In Section 3., we describe the data that we used to build the models and what material we test on. The experiments are described in Section 4., and we draw conclusions in Section 5..

2. Preliminaries

In this section, we briefly list the software and hardware used in the experiments, and describe the mechanisms of the TETRA codec.

2.1. Software and Hardware

For feature extraction, we employ the HTK toolkit (Young, 1994), and extract 39 features (12 MFCCs, Energy, Delta, Acceleration and Zero Mean) for each frame of 25 ms window length, using a stepsize of 10 ms. For language modeling, we make use of the MIT Language modelling toolkit (Hsu and Glass, 2008) to compute a trigram language model with modified Kneser-Ney smoothing. We use the Julius toolkit (Lee et al., 2001) for decoding.
We employ the CM 5000 radio station and the MTP 850 handheld device, both by Motorola (see Figure 1).

2.2. TETRA

Terrestrial trunked radio (TETRA) is a standard for a digital trunked radio system. It was published in the mid 90s by the European Telecommunications Standards Institute (ETSI). 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 a_i linear prediction coefficients of order $p = 10$, the short-term synthesis filter is given by:

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

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

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

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:

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

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 $A(z)$, given by the Toeplitz lower triangular matrix that is constructed from the filter impulse response:

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

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 (ETSI, 1998). See Figure 2 for an example of the channel effect on a single word, based on the frequency analysis.

3. Data

In this section, we describe the corpora used to build the models, and introduce the PRONTO corpus as our target material for recognition. Our initial models, as pointed out below, were taken from a news domain and thus required some effort for adaptation.

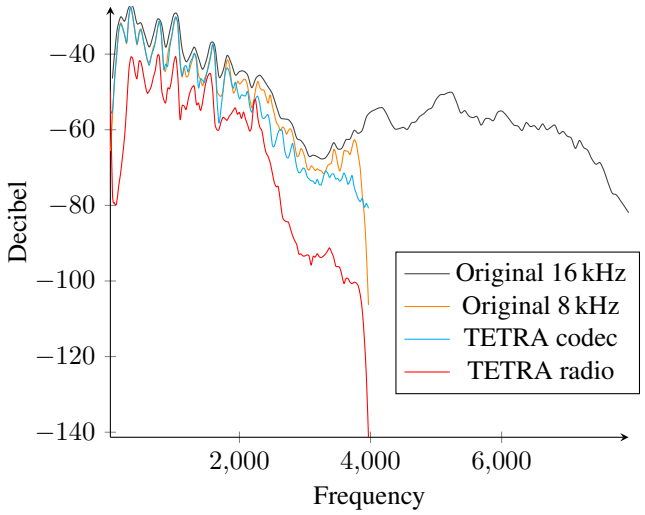

Figure 2: Frequency analysis on the word “Tagesthemen”

3.1. Acoustic Model Training Material

For training of the acoustic model, we employ 82 799 sentences from transcribed video files. They are taken from the domain of broadcast news and political talk shows. The audio is sampled at 16 kHz and can be considered to be of clean quality. Parts of the talk shows are omitted when e.g. many speakers talk simultaneously or music is played in the background.

3.2. Language Model Training Material

The language model consists of the transcriptions of the audio files as described above, plus additional in-domain data taken from online newspapers and RSS feeds. In total, the material consists of 11 670 856 sentences and 187 042 225 running words. Of these, the individual subtopics were used to train trigrams with modified Kneser-Ney discounting, and then interpolated and optimized for perplexity on a with-held 1% proportion of the corpus.

3.3. PRONTO Corpus

The firefighter data was collected as part of the EU-funded project “Event Recognition for Intelligent Resource Management” (PRONTO)[1].
In total, the broadcast of ten firefighter exercises have been recorded. Figure 3 shows a picture taken at one of the emergency scenarios, where a fire spread in the first floor of a building used for the exercises. The material consists of status reports on the place of accident, conveyance of contaminant analysis, request for backup, et cetera. A portion of the material has been recorded up to two times, but on different broadcast stations which differ highly in their speech quality. Other sentences are more common (for example, when the fire engine calls the operator station), but are uttered by different speakers. We had no access to which sentence was spoken by which speaker, since they were all sharing the same communication infrastructure. Based on the inital set of 1 769 transcribed sentences, 1272 (71.9%) sentences

[1] http://www.ict-pronto.org/

(a) CM 5000 radio station used for broadcast

(b) MTP 850 handheld device receiving the signal

Figure 1: Motorola equipment used for the recording and transmission of audio signals. Both devices use the TETRA encoding scheme for internal communication.

Figure 3: sample emergency exercise scenario

Table 1: Corpus statistics for the PRONTO corpus. Of all the words that are out-of vocabulary (OOV), 98 occur in both sets, and 103 occur only in the test set.

	dev	test
# sentences	769	1 000
# running words	5 548	7 235
distinct words	810	982
OOVs (as per phon. dictionary)	166	201
running time [h]	0:40	0:54

are unique. The data has been randomly split into 769 sentences for the development set and 1000 sentences for the test set.
For a complete overview of the corpus statistics, see Table 1.

3.3.1. Audio Analysis

The data has been recorded under real-life conditions. Slip of the tongue happens occasionally, hesitations occur frequently. Some parts are recorded in-door, others on the street. Background noise occurs frequently, occasionally there are co-interference phenomena from different channels or mobile phones. Sirens from emergency vehicles are audible in several instances. A few utterances are not intelligible – in these instances, the communication partner usually asks the speaker to repeat the statement. Whenever we can deduce the original utterance from this dialogue, we transcribe the most-likely sentence for the unintelligible part. Since a radio button on the handheld device has to be pressed before speech is recorded, the beginning and the end of an utterance is often truncated. Parts of the material, especially places and numbers, are spoken with a local dialect. While we can estimate the influence of the TETRA codec, the signal filter induced by the main radio station and the handheld device are unknown. Although the data has been sampled at 16 kHz, the TETRA codec itself already acts as a low-pass filter for frequencies below 4 kHz.
To obtain a first feeling for the expected performance loss through the hardware setting, we carried out preliminary experiments on with-held data taken from 5 719 sentences of in-domain news data. The clean training material has been transformed in various ways, whereafter the acoustic model was retrained from scratch. The recognition performance was finally measured on a simultaneously transformed in-domain test set, in order to estimate the expected performance degradation due to the channel.
See Table 2 for the results. By downsampling, the WER performance dropped by 2.5% absolute, and the TETRA codec already results in 10.8% WER loss absolute. By employing the broadcast station on both training and test material, we see an additional performance drop by 4.9% WER absolute, for a total 25.6 WER degradation.
In general, we feel that the TETRA encoding scheme introduces challenging distortion onto the speech signal. Apart from acting as a low-pass filter, preliminary findings in (Stein et al., 2012) suggest that the codec emphasizes harmonic distortions of the audio hardware, due to its built-in

Table 2: Expected performance loss through TETRA codecs, tested on in-domain news data of withheld 5 719 sentences. Same preprocessing of training & test material.

transformation	WER
clean 16 kHz data	26.6
downsampling 8 kHz	29.1
+ TETRA codec	37.4
TETRA broadcast station	42.3

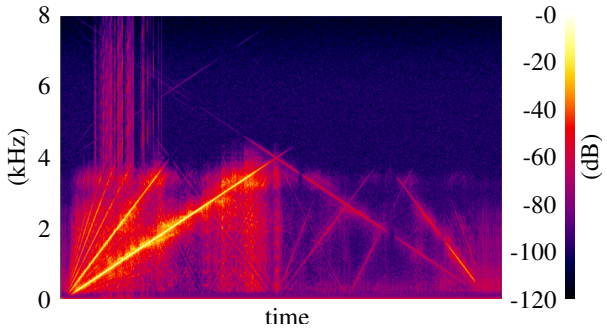

Figure 4: Distortion of the signal as introduced by the TETRA codec and the hardware equipment: spectrogram of an 8 kHz sweep as received over a TETRA radio station

adaptive code book which has been optimized on human speech intelligibility. While for a human ear, a total harmonic distortion of up to 10% can be inaudible (Geddes, 2003), the automatic speech recognition is severely hampered; this becomes most apparent in substitution errors for phonemes with pronounced overtones. See Figure 4 for the distortion introduced when transmitting a sweep signal.

3.3.2. Domain Analysis

The material consists of many firefighter termini. Longer words that are common are often abbreviated. Due to the two-way radio systems, voice procedure (e.g. "affirmative", "over and out") is used very frequently. The grammar is often quite basic, verbs are often in infinite form.

When computing the trigram perplexity on the firefighter domain, the news material is obviously a bad starting point, being more than ten times higher on the PRONTO sets than for in-domain data. This is not only due to the different vocabulary: to demonstrate the grammar effect on the part-of-speech level, we parsed the whole training sentences used for the language model as well as the PRONTO corpus using TreeTagger,[2] and computed the trigram perplexity of a POS-based language model on the development set. Even here, the perplexity nearly doubles. See Table 3 for an overview of the results.

Even though a language model on the development set has the best results, we opted for a linearly interpolated language model of the news data and the development set. Otherwise, all words not apparent in the development data would be mapped to "unknown", and the models would thus be rendered not very stable beyond the current sessions. We will discuss this detail by adding more in-domain data in a later section.

[2] http://www.ims.uni-stuttgart.de/projekte

Table 3: Domain analysis on the PRONTO corpus, via trigram language model perplexity.

	trigram perplexity		
	news	dev	test
news	170.4	2387.5	2263.0
POS news LM	8.0	15.7	16.3
dev		5.3	25.2
news + dev, linear interpolated		10.1	56.4

4. Experiments

Based on the analysis conducted in the previous section, we expect the following error sources. First, there is an obvious mismatch between the clean acoustic material of broadcast news and the one taken from the firefighter scenario. Also, the additional written material used for training the language model is of quite a different nature. We try to reach a first reasonable baseline with standard adaption and interpolation techniques (Sec 4.1.). Second, several domain-specific problems like dialect and voice procedure grammar seem to arise. Here, we can employ knowledge of the material to extend the dictionary and the language model (Sec 4.2.). Third, we lack sufficiently sized training material. We semi-automatically expand our material by transcribing new audio sets, and estimate the influence of the new transcriptions on the recognition quality after every iteration. Further, we extend the written material by crawled in-domain data (Sec 4.3.).

4.1. Baseline Model Adaptation

Since the acoustic signal recorded in the sessions had the same sampling frequency as our acoustic speech recogntion model, 16 kHz, we ran a first experiment simply reusing our ASR architecture. As expected, this approach does not produce any valid output, the error rate is practically 100%. Maximum A Posteriori (MAP) adaptation of the mixture models on the features extracted from the development set heighten the performance to a meager 83.2% WER. This was obviously due to the mismatch in the actual frequency range, since TETRA acts as a low-pass filter. In a second line of experiments, we resampled the acoustic training (i.e. the news broadcast) to 8 kHz and retrained the acoustic models. This raised the word accuracy by 9.1% absolute. Finally, a simple linear interpolation of a language model trained on the development set and the news language model gave a major boost and resulted in a total WER of 51.8%. This illustrates a high similarity in the development and the test set, but as stated before, same sentences are spoken under different audio conditions and by different speakers. See Table 4 for the results.

Table 4: Performance on the PRONTO corpus for the various adaptation procedures.

	WER
base, no adaption	98.5
+ MAP adaptation	83.2
+ AM 8 kHz resampling	72.1
+ LM interpolation	51.8

Table 5: Example for dialectal adaptation of the pronunciation dictionary.

	fünfundneunzig (59)
High German	f Y n f Q U n t n OY n t s I C
Dialect	f Y m @ n n OY n t s I S

4.2. Domain-specific Experiments

As noticed in the audio analysis, most of the firefighters speak a local dialect and pronunciation variant. Especially in the dialogue parts which have a repetitive structure the dialect tends to be quite thick. Since this typically involves names and lots of numbers, we opted to adjust the pronunciation dictionary for these entities. See Table 5 for an example.

A bad phoneme dictionary not only affects the recognition, but also the linear alignment step during the MAP adaptation, and correcting these gave an improvement of 1.9 absolute.

In a second step, we employed a post-processing step which concatenated numbers that were clearly forming a unity but were recognized as separate (e.g. “eighty one”). Since these are written as compound words in German, they were recognized as errors before, and fixing this gave another 0.9 improvement.

In a last domain-specific step, we enlarged the language model derived from the development set as follows: we removed any hesitations and added both sentences. Then, we noted all voice procedure words at the end of sentences (e.g. “affirmative”, “come in”) and added seven possible marker words to each and every sentence, replacing the existing ones if apparent. In a last step, we looked for occurrences of “Florian” (the patron saint of the firefighters) and assumed every following word with a capital letter to be the location. Then, we replaced this location with every other possible location. Since each of these steps accumulated the sentences that were generated, we ended up with 16 477 sentences derived from 769 sentences, and call this approach *bloated LM*. It improved the WER by another 0.9 points absolute. See Table 6 for the results.

4.3. Semi-automatic Corpus Expansion

Due to the limited size of the transcribed material, we face a tuning problem. While a simplified solution like linear interpolation of a 11 M out-of-domain corpus with a 769 development corpus for language model creation is unsatisfactory, there is no further data for a weighted interpolation without losing substantial proportions of the material. Also, tuning of e.g. language insertion penalties and noise threshold suffers from the same conditions. However, we still have unannotated material, and at this stage the recognizer seems already good enough to support the transcription procedure.

We opted for a semi-automatic corpus extension as follows: we join the development and the test set for a new, bigger training set, and automatically segment the new data into sentence chunks. We then run the recognizer on the unseen data, manually correct 200 sentences, compute their error and add this material to the training set. Note that, since all material is hand-corrected regardless of the recognition output, we still can opt to employ the new material and recognize on the test set in a later step (whilst removing it from the training material, of course).

The idea is that it should become increasingly easy to annotate new material since the recognizer should become more reliable after each step. Also, we should be able to see the performance gain after each step and can decide whenever we consider the trade-off between new annotation time and performance gain is still high enough. We plotted the progress of the recognizer on unseen data in Figure 5.

Adding the newly annotated date to our original setup (where the test set is excluded from the training material), this yields an improvement of 1.8% WER on the best sys-

Table 6: Performance on the PRONTO corpus for the domain-specific enhancements.

	WER
adapted model	51.8
+ dialectal pronounciation	49.7
+ post-processing of numbers	48.8
+ bloated development LM	47.9

Figure 5: Performance gain through corpus extension

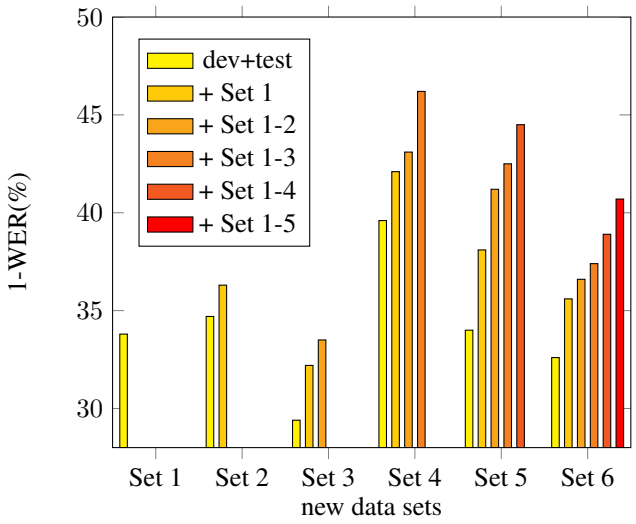

Table 7: Performance on the PRONTO corpus.

	IdF test set WER
adapted + enhanced	47.9
+ additional ac. material	46.1
+ additional text material	45.2

tem.

Next, we also extended the in-domain written text collection by crawling firefighter websites for operational reports, for a total of seven cities. This resulted in 30 791 running sentences, containing 318 954 words. The dictionary has been extended accordingly, leading to an improvement of 0.9% WER. See Table 7 for an overview of the results.

5. Conclusion

While both the material and the domain of firefighter radio transmission is challenging, standard methodology already leads to promising results. With suitably tailored domain-specific enhancements, the recognizer is soon at a stage where it can substantially support further annotation procedure, so that within reasonable time new data can be acquired.

Manually checking the recognizer output, especially the standardized voice procedures that request attention of one party for the other, and the standardized calls for backup, are among the sentences that perform best in accuracy. Weaker sentences include those that are specific to the situation, like e.g. the exact nature of the current emergency situation, and unforeseen issues like locked doors or leaking chemical barrels. We therefore conclude that the recognizer can already be utilized in a real-life scenario, by e.g. logging communications, and suggesting backup, so that the human interaction can focus on a flexible response to the specific situation.

Acknowledgments

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.

6. References

ETSI. 1998. Terrestrial Trunked Radio (TETRA); Speech Codec for Full-rate Traffic Channel; Part 2: TETRA Codec. Technical Report ETS 300 395-2, European Telecommunication Standard, February.

S. Euler and J. Zinke. 1994. The Influence of Speech Coding Algorithms on Automatic Speech Recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume i, pages I/621–I/624, April.

Lidia W. Geddes, Earl R.; Lee. 2003. Auditory perception of nonlinear distortion - theory. In *Audio Engineering Society Convention 115*, 10.

Bo-June Hsu and James Glass. 2008. Iterative Language Model Estimation: Efficient Data Structure & Algorithms. In *Proc. Interspeech*.

Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. 2001. Julius — an Open Source Real-time Large Vocabulary Recognition Engine. In *Proc. of the EUROSPEECH*, pages 1691–1694.

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

C.H. Slump, T.IJ.A. Simons, and K.A.Verweij. 1999. On the Objective Speech Quality of TETRA. In *Proc. of the Anual workshop on Circuits, Systems and Signal Processing*, pages 421–429, Mierlo, the Netherlands, November.

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

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

S.J. Young. 1994. The HTK Hidden Markov Model Toolkit: Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44.