Speech Recognition with μ-Law Companded Features on Reverberated Signals

Tino Haderlein, Georg Stemmer, and Elmar Nöth

University of Erlangen-Nuremberg, Chair for Pattern Recognition
(Informatik 5), Martensstr. 3, 91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de,
http://www5.informatik.uni-erlangen.de

Abstract. One of the goals of the EMBASSI[1] project is the creation of a speech interface between a user and a TV set or VCR. The interface should allow spontaneous speech recorded by microphones far away from the speaker. This paper describes experiments evaluating the robustness of a speech recognizer against reverberation. For this purpose a speech corpus was recorded with several different distortion types under real-life conditions. On these data the recognition results for reverberated signals using μ-law companded features were compared to an MFCC baseline system. Trained with clear speech, the word accuracy for the μ-law features on highly reverberated signals was 3 percent points better than the baseline result.

1 Introduction

One of the major goals of the EMBASSI project is to develop human-machine interfaces for television sets and VCRs. The user's speech is supposed to replace a conventional remote control. As it is inconvenient to learn a fixed set of instructions the devices will have to understand spontaneous speech. Linguistic analysis and speech understanding are therefore large working areas in the project. Others are signal enhancement and speech recognition in a reverberated environment.
The user's utterances would be optimally received by a close-talk microphone. This would mean, however, that you have to wear a headset while watching TV or speak into a hand-held microphone. It is obvious that the vast majority of consumers would not accept this. The microphones will rather have to be integrated into the device itself or distributed within the room. On the long way from the speaker to the microphone(s) many different kinds of distortions may influence the signal:

- reverberation from the surrounding walls and windows
- talk of other persons in the room
- background music or running TV program

[1] http://www.embassi.de

published in: Text, Speech and Dialogue (TSD 2003), Proceedings, LNAI 2807, Springer, Berlin, Heidelberg, 2003, pp. 173-180

– other types of noise, e.g. from outside
– varying room acoustics, e.g. by opening the door or moving persons

This paper concentrates on the problem of developing features suitable for reverberated signals. An overview of environment-independent features and recognition is given in [1, pp. 39-51] and [2]. It includes features like the Root Cepstrum Coefficients (RCC, [3]) which seem to be less affected by additive background noise than MFCCs [4]. Perceptual Linear Prediction (PLP, [5]) also contains as one computation step an auditory-like cubic root compression. PLP features, combined with several RASTA filtering methods, were successfully applied on data with additive and convolutional noise [6, 7].
Recent research shows that reverberant speech recognition is improved by long-term spectral subtraction [8]. If synchronously recorded data from close-talk and distant microphones is available, Neural Networks can be trained e.g. in the cepstral domain to transform a reverberated signal into its non-reverberated counterpart in order to compensate the distortion [9].
Furthermore good results in hands-free speech recognition have been achieved by combining the signals from a microphone array [10]. This is also planned in the EMBASSI project.

2 The EMBASSI Speech Corpus

In order to work with realistic data a German speech corpus was recorded which contains most of the mentioned influences. Recordings were made in a room which was in its acoustical properties equal to a living-room. All walls of the room were equipped with a curtain which resulted in a reduced reverberation time of $T_{60} = 150$ milliseconds, which means the time span during that the reverberation decreases by 60 dB. In this room recordings were made with 20 speakers (10 male, 10 female) who were between 19 and 29 years old.
A close-talk microphone (headset) and an array of 11 microphones were used. The array microphones were mounted in one line in a height of 116 cm in front of the speakers. The two microphones to the very left and right had a distance of 16 cm to their neighbours, all the others were 8 cm apart from each other.
Experiments in an early phase of the EMBASSI project had shown how people would talk to a TV set or a VCR, if speech input were supported. Taking these into account sentence templates were modeled and an automatic text generator produced the sentences to be read by the speakers. They consisted of commands like e.g. "*I'd like to see '<TV show>' please.*", "*Turn up the volume.*" or "*What is running at <time> on <channel>?*".

The recordings were divided into two blocks:
1. The first block included a disturbing speaker and noise from loudspeakers in the room. For each speaker four sessions were made: 1. without further background noise, 2. with rock music at a moderate volume, 3. with loud rock music and 4. with a "newsreader" (actually an interpreter recorded in her booth). One session lasted about 90 seconds, the speaker read 26 sentences. The distance between speaker and microphone array was 1 meter.

2. As it was not possible to record all possible kinds of noise together with the speaker, the idea was to record the undistorted speech and mix noise to these signals later. However, the signals achieved with this method will not be equal to signal from a real-life situation. One important aspect is that the Lombard effect is missing. This means that a person's voice will get higher and louder in noisy environment. Therefore the mentioned noise types were given on the speaker's headphones instead of on the loudspeakers in the room and the speaker's clear voice with Lombard effect was recorded. No other persons were present in the room during these experiments. The distance to the microphone array was either 1 meter or 2.5 meters. For both distances five sessions were made, where no. 1 to 4 were the same as above and in no. 5 the "newsreader" was played onto the headphones loudly. One session lasted between approx. 150 and 180 seconds, the speaker read 60 sentences.

The 20 persons read a total of 15360 commands. The total duration of signals recorded is about 11 hours. The data were recorded in CD quality (48 kHz sample frequency, quantized at 16 bit). For the experiments described in the following sections the data were downsampled to 16 kHz.

3 The Baseline System

3.1 Data and Feature Set

All experiments in this paper are based on the recordings of the EMBASSI corpus where the speaker was alone and it was silent in the room so that no Lombard effect occurred. Unfortunately these were only about 100 minutes of speech, but testing a new feature set also means an entire recognizer training each time, so a compromise had to be made between robustness and performance. The training data consisted of the close-talk recordings of 6 male and 6 female speakers (60 min of speech, 8315 words). One male and one female speaker were the validation set (10 min, 1439 words), and one half of the test set consisted of the remaining three men and three women (30 min, 4184 words). The other half were the corresponding data of the central array microphone, which was 1 m away during one of the used sessions and 2.5 m during the other. In order to optimize the training the session files were cut into pieces containing one single sentence each.

In all experiments the number of features was kept at 24. The features for the baseline system were the signal energy, 11 MFCCs and the first derivatives of those 12 static features. The derivatives were approximated by the slope of a linear regression line over 5 consecutive frames (50 ms). In all experiments only the compression function was changed.

3.2 Experiments with the Baseline System

Our speech recognition system uses semi-continuous HMMs. It models phones in a context as large as still statistically useful and thus forms the so-called polyphones. The HMMs for each polyphone have three to four states. The recognizer

Table 1. Word accuracies (WA) for the baseline system for three microphone distances

microphone distance	language model	word accuracy
close-talk	4-gram	94.2
close-talk	0-gram	69.8
1 m	4-gram	89.6
1 m	0-gram	52.1
2.5 m	4-gram	82.2
2.5 m	0-gram	36.6

has a vocabulary size of 474 words and was initially trained with a 4-gram language model. This baseline system achieves a word accuracy (WA) of 94.2% on the close-talk recordings, 89.6% for the array microphone recordings at a distance of 1 m and 82.2% for the 2.5 m microphone distance. When the language model was switched off ("0-gram model") the close-talk word accuracy dropped to 69.8%. For the 1 m distance 52.1% were reached, for the 2.5 m distance 36.6% (compare Table 1). As can easily be seen the word accuracy is significantly lower when the distance between the speaker and the microphone grows. In the following we will investigate different feature sets in order to reduce this decrease in performance.

4 Alternatives to the Mel-Cepstrum

4.1 Motivation

The problem with the logarithmic compression of the filterbank coefficients is that it is most sensitive to spectral parts with the lowest power, i.e. where the signal-to-noise ratio (SNR) is usually worst [2]. Furthermore low feature or coefficient values below 1 can cause problems with the float number range of the computer. Solutions for this problem can be to replace $\log(x)$ by $\log(x+c)$, where c is a small constant or the introduction of a minimum threshold to which critical values will be set. Alternatively the log function can be omitted at all in favor of functions with more suitable companding characteristics, e.g. with root functions computing the root cepstrum as introduced in [3]. It simply replaces the logarithm by a root function $\sqrt[n]{x}$.
The μ-law (or "mu-law") coding has the formula

$$f(x) = x_{max} \cdot \text{sign } x \cdot \frac{\log(1+\mu|x|/x_{max})}{\log(1+\mu)} \quad \text{where sign } x = \begin{cases} +1 & \text{for x} > 0, \\ 0 & \text{for x} = 0, \\ -1 & \text{for x} < 0. \end{cases}$$

In our approach x_{max} was equal to 1, because before the companding an energy normalization was made. Fig. 1 show examples for several values of μ. Like a root function the μ-law functions raise low values and scale down, depending on

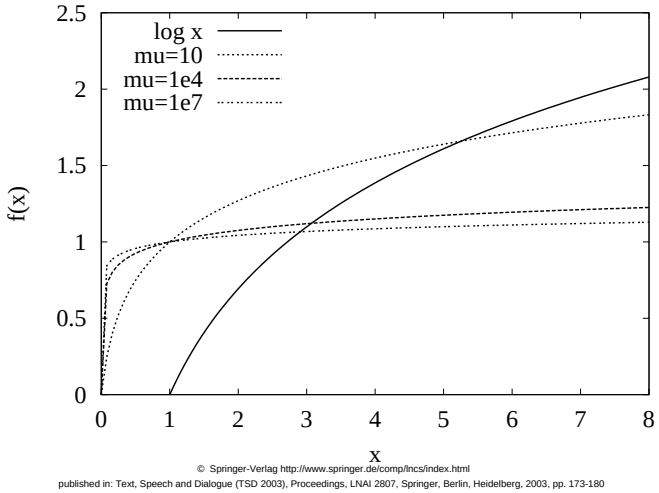

Fig. 1. The $\log_{10}$ function and some μ-law characteristics

their parameter, high values stronger than logarithmic compression. A similar idea has also been used within the RASTA methodology, when in J-RASTA the logarithm before the filtering was replaced by $\log(1+JX_i)$, where i is the critical band number and J is a user-defined factor [6, 11].

4.2 Applying μ-Law Compression

Several recognizer trainings were performed on close-talk recordings, this time using the μ-law compression during the feature extraction. Powers of 10 were set as values for μ. The recognition results are graphically presented in Fig. 2 for the enabled language model and in Fig. 3 for the case without linguistic help. The word accuracies are also summarized for the best μ values in Table 2. A slight improvement of the recognition could be achieved for all experimental settings, i.e. for all three microphone distances and for enabled and disabled language model. The close-talk signals reached the best word accuracy at the highest tested value $\mu = 10^9$ with the language model (94.8%, baseline: 94.2%) and also without it (70.7%, baseline: 69.8%). Thus even higher values for μ will have to be tested. The recordings with 1 m microphone distance were recognized best at $\mu = 10^6$ with the language model (91.4%, baseline: 89.6%) and $\mu = 10^7$ without a language model (53.1% vs. 52.2%). The 2.5 m distance recordings had their maximum at $\mu = 10^5$ (83.3% vs. 82.2%) and $\mu = 10^6$ (39.6% vs. 36.6%), resp. It seems that μ should be the smaller the higher the reverberation in the signal is.

It was reported for Root Cepstrum Coefficients that, at least for Linear Frequency Cepstral Coefficients (LFCC), different root functions for training and test set can improve performance if noisy signals are tested on a recognizer

Table 2. Word accuracies for features using μ-law compression. All three microphone distances were tested with and without a language model

microphone distance	language model	word accuracy μ=10^4	μ=10^5	μ=10^6	μ=10^7	μ=10^8	μ=10^9	baseline
close-talk	4-gram	94.5	94.7	94.5	94.6	94.6	94.8	94.2
close-talk	0-gram	69.6	70.1	69.7	70.3	69.8	70.7	69.8
1 m	4-gram	89.7	91.1	91.4	90.8	90.3	90.3	89.6
1 m	0-gram	51.7	52.6	52.3	53.1	51.8	52.1	52.1
2.5 m	4-gram	80.7	83.3	83.0	81.0	81.4	81.1	82.2
2.5 m	0-gram	35.8	37.8	39.6	37.5	36.5	37.0	36.6

trained with clear speech [12, 13]. In order to find out if this might also be valid for μ-law companded features, the 2.5 m microphone distance data were tested on the $\mu = 10^9$ recognizer with some μ values smaller than the one for training. But compared with the original 81.1% and the baseline 82.2% the word accuracy reached only about 75% for several steps between $\mu = 9 \cdot 10^8$ and $\mu = 10^9$ and dropped further for lower parameters. This work is still in progress.

5 Conclusions and Outlook

In this paper the Mel-Frequency Cepstrum Coefficients (MFCC) were compared to cepstral coefficients which had been μ-law companded after the spectral filterbank. The test data were speech signals recorded in quiet environment with a close-talk microphone and a microphone at a distance of 1 m or 2.5 m to the speaker. Speech recognizers were trained with 60 minutes of close-talk recorded speech and tested on signals of all three microphone distances. The μ-law companded features managed to outperform the baseline system by about 1% word accuracy (WA) absolute for each of the test cases and reached even at 2.5 m microphone distance 83.3% WA (baseline: 82.2%). The improvement relative to the word error rate was with an active 4-gram language model 10.3% for the close-talk signals, 17.3% for 1 m microphone distance and 6.2% for the 2.5 m distance. The corresponding values for the inactive model were 3.0%, 2.1% and 4.7%, resp. For reverberated signals a lower μ parameter seems to be better than for undistorted signals. Further experiments have to be made to examine this.

In current test series not only the first derivative of the features is taken into account, but also the second. In the future the number of features (here: 24) will be varied and combined with decorrelation methods like PCA. Neural Networks for transforming distorted into undistorted features, as described in [9], will also be applied. Basic work for this has already been done [14].
Furthermore the MFCCs with logarithmic companding have to be compared to PLP features which produced good results on reverberated data [6, 11].

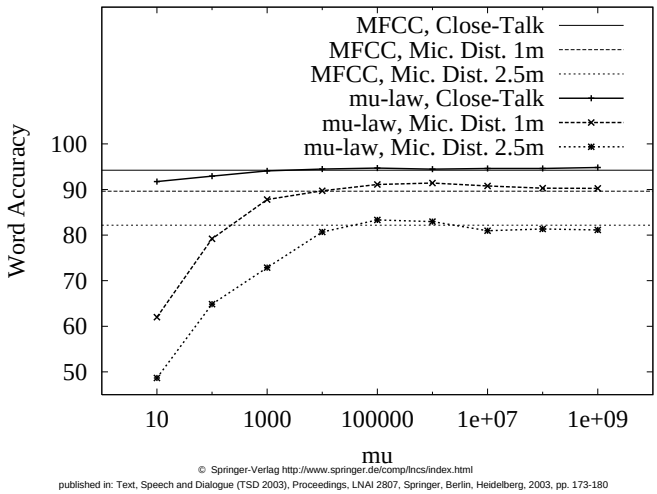

Fig. 2. Word accuracies for the recognizers using μ-law features; all three microphone distances were tested including a 4-gram language model. The straight lines show the results for the baseline system

Acknowledgments

The EMBASSI project was supported by the German Federal Ministry of Education and Research (grant no. 01 IL 904). The responsibility for the contents of this study lies with the authors. The speech corpus used in the described experiments was recorded at the Chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg (Department of Electrical, Electronic and Communications Engineering).

References

[1] Junqua J.-C.: Robust Speech Recognition in Embedded Systems and PC Applications. Kluwer Academic Publishers, Boston (2001)
[2] Hunt M.J.: Spectral Signal Processing for ASR. In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Vol. 1. Keystone, Colorado (1999) 17–25
[3] Lim J.S.: Spectral Root Homomorphic Deconvolution System. IEEE Trans. ASSP, Vol. 27. 3 (1979) 223–233
[4] Sarikaya R., Hansen J.H.L.: Analysis of the Root-Cepstrum for Acoustic Modeling and Fast Decoding in Speech Recognition. In Proc. European Conf. on Speech Communication and Technology (Eurospeech), Vol. 1. Aalborg, Denmark (2001) 687–690
[5] Hermansky H.: Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of The Acoustical Society of America, Vol. 87. 4 (1990) 1738–1752
[6] Koehler J., Morgan N., Hermansky H., Hirsch H.G., Tong G.: Integrating RASTA-PLP into Speech Recognition. In Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Vol. 1. Adelaide, Australia (1994) 421–424

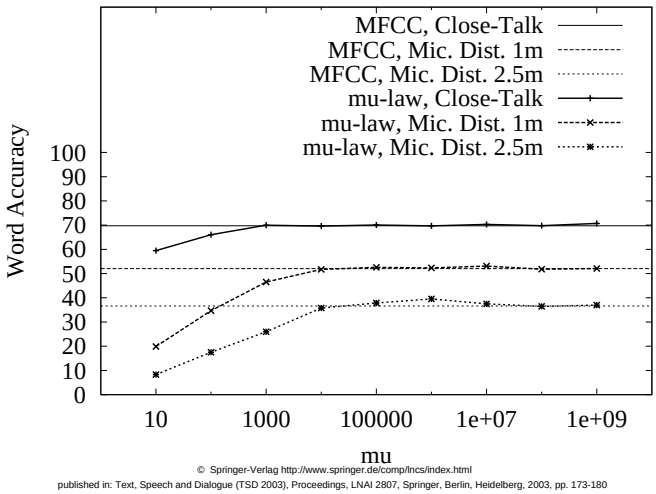

Fig. 3. Word accuracies for the recognizers using μ-law features; all three microphone distances were tested and the language model was inactive. The straight lines show the results for the baseline system

[7] Kingsbury B.E.D., Morgan N.: Recognizing Reverberant Speech with RASTA-PLP. In Proc. Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Vol. 2. Munich, Germany (1997) 1259–1262

[8] Gelbart D., Morgan N.: Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition. In Proc. Int. Conf. on Spoken Language Processing (ICSLP), Vol. 3. Denver, Colorado (2002) 2185–2188

[9] Pan Y., Waibel A.: The Effects of Room Acoustics on MFCC Speech Parameter. In Proc. Int. Conf. on Spoken Language Processing (ICSLP), Vol. IV. Beijing, China (2000) 129–133

[10] Omologo M., Svaizer P., Matassoni M.: Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Communication, Vol. 25. 1–3 (1998) 75–95

[11] Morgan N., Hermansky H.: RASTA Extensions: Robustness to Additive and Convolutional Noise. In Proc. Workshop on Speech Processing in Adverse Conditions. Cannes, France (1992)

[12] Alexandre P., Lockwood P.: Root cepstral analysis: A unified view. Application to speech processing in car noise environments. Speech Communication, Vol. 12. 3 (1993) 277–288

[13] Lockwood P., Alexandre P.: Root Adaptive Homomorphic Deconvolution Schemes for Speech Recognition in Noise. In Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Vol. 1. Adelaide, Australia (1994) 441–444

[14] Weiß R.: Anwendung von KNN zur Beseitigung der raumbedingten Störungen in einem Sprachsignal. Student Thesis (in German), Chair for Pattern Recognition, University of Erlangen-Nuremberg (2002)