An Improvement of robustness to speech loudness change for an ASR system based on LC-RC features

Pavel Yurkov, Maxim Korenevsky, Kirill Levin
Speech Technology Center, St. Petersburg, Russia

Abstract

This paper deals with new front-end feature improvements for Automatic Speech Recognition (ASR) robustness to changes in speech loudness. Our experiments show that applying a RASTA–like filter gives a significant improvement in robustness to speech loudness change, leading to an up to 4% PER reduction.

Keywords: long-term features, RASTA, ANN, LC-RC features, ASR robustness.

1. Introduction

TRAP–like long-term features used for ASR give a significant reduction of error [1-3]. A further long-term front-end development is the use of LC-RC features [4, 5]. In this paper we present the research results of LC-RC feature robustness. A number of improvements are proposed.

This paper is organized as follows. Section 2 of the paper describes the speech corpus used for the experiments. Section 3 describes LC-RC features and their advantages and drawbacks. Two kinds of LC-RC feature enhancements that increase robustness to loudness change are proposed in Sections 4 and 5. These enhancements are experimentally checked and the results are presented in Section 6. Section 7 contains an analysis of the results.

2. The speech corpus

We used the Russian SpeechDat(E) corpus. It was divided into training (about 30 hours, 1572 speakers), cross-validation (about 4 hours, 210 speakers) and test (about 6 hours, 314 speakers) sets. Each set contains an equal number of male and female voices.

3. Experimental setup

Fig. 1 shows the structure of the baseline LC-RC system. At the first stage, each 10-ms Mel filter-bank output is merged into a super vector 310 ms in length (31 frames). Then the super-vector is split into left (1-16 frames) and right (16-31) time contexts. Both parts are weighted by the respective half of a Hamming window and DCT is applied. We use C_0-C_{11} cosine coefficients.

Then the left and right super-vectors are used as the input of two MLP classifiers. The output posteriors of LC and RC MLPs are log-weighted and sent to the third MLP (merger). The merger returns posterior probabilities of phonemes which are passed to a Viterby decoder to find the most probable phoneme sequence.

Left and right context MLPs have the same topology: 240 inputs in the first layer, 1000 neurons in the hidden layer and 163 neurons in the output layer. The merger MLP also has three layers, but a different number of inputs in the first layer.

We have used a tandem scheme of ASR where posteriors of the merger are used as front-end vectors for the decoder.

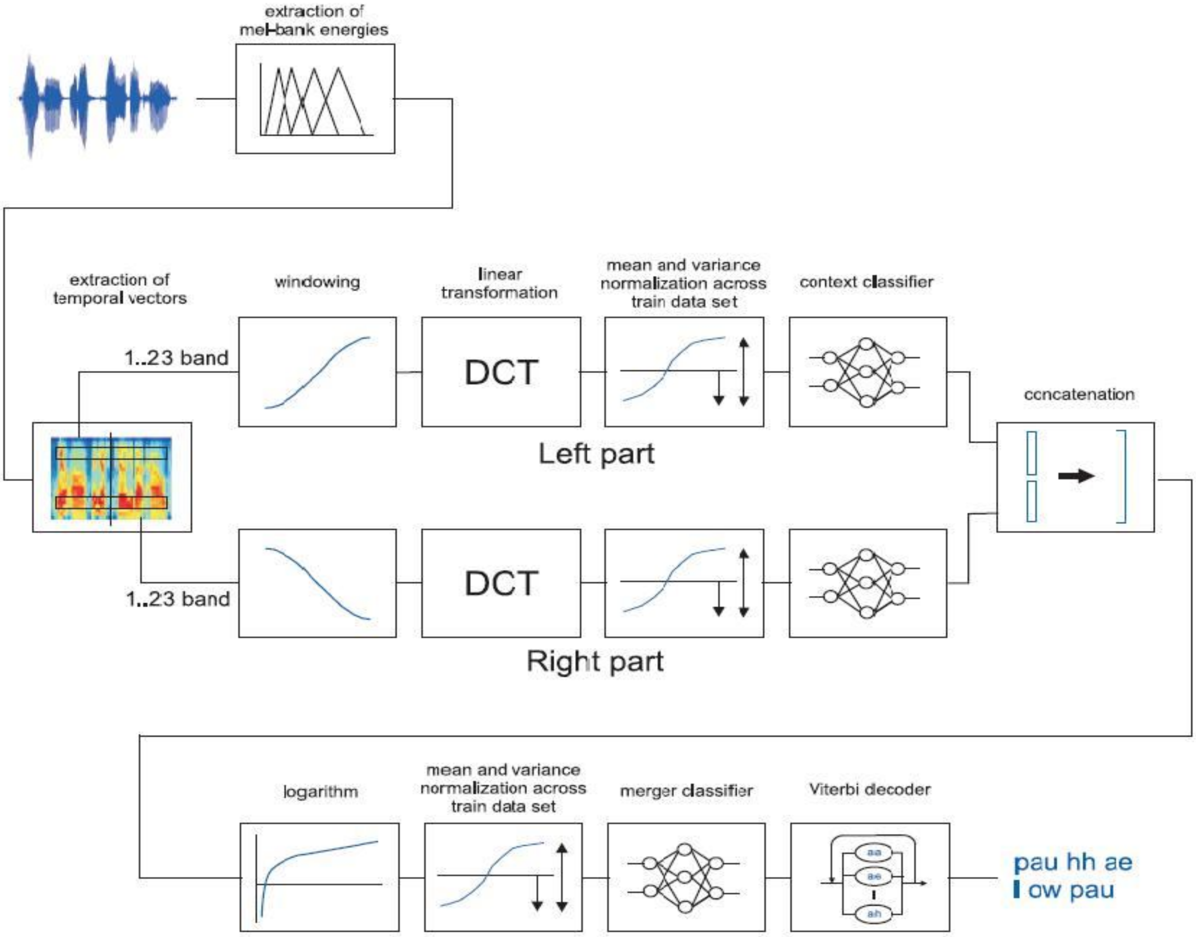

Fig.1 The baseline LC-RC ASR architecture [4]

The baseline system uses C_0 of the DCT transform because it gives valuable information about speech energy and increases recognition accuracy (about 2-3%). The main drawback of C_0 usage is that it makes the ASR sensitive to speech loudness. We propose two solutions for eliminating this drawback.

4. C_0 normalization

Vocalized sounds are defined by their spectral maxima (formants). As we mentioned above, since C_0 contains information on the energy of the signal, this means that the set of C_0 coefficients of all critical bands at time T describes formants. C_0 deletion leads to formant positions loss.

The fist method of LC-RC enhancement that we propose substitutes C_0 of every critical band according to the following equation:

$$C_0(i) = C_0(i) - C_0(1), \quad i = 1..n,$$

where i is the index of the corresponding critical band. This kind of C_0 representation, on the one hand, saves information on spectral maxima, and on the other hand minimizes dependency on speech loudness.

5. RASTA-like filtering

The second method is based on the same idea as the cepstral mean normalization (CMN) procedure. Unfortunately, canonical CMN requires a relatively wide time window to estimate the mean so it is applicable only for off-line tasks. For on-line speech processing a fixed width window can be used or exponential smoothing can be applied. The correlation time of the filter must correspond to the syllabic frequency of the vocal tract. This type of filtering (RASTA) was proposed by Hermasky [6, 7]. Figure 2 shows the frequency response of our RASTA-like filter. Its characteristics slightly differ from the original RASTA, which gives a significant performance improvement [8].

The filter smoothes high frequencies and amplifies syllabic frequencies at 4-6 Hz which emphasize the speech-specific characteristics of the audio signal.

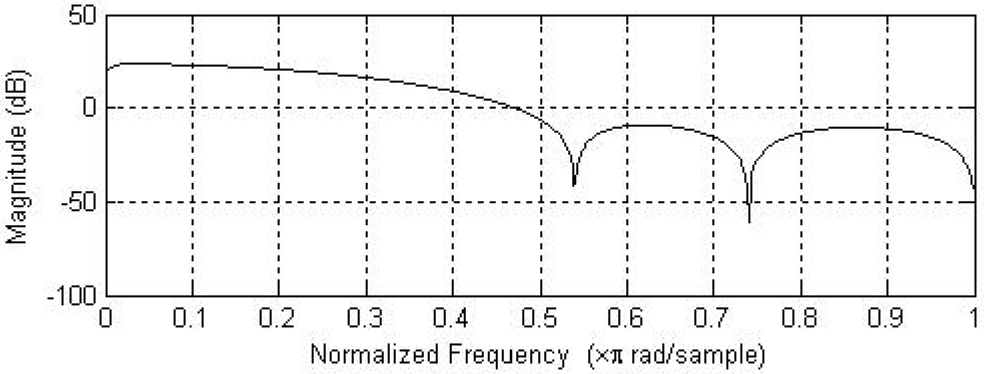

Fig. 2 Frequency response of RASTA-like filter

6. Experiments and Results

We have trained three types of ASR system corresponding to baseline, RASTA-like and C_0 normalized features. We do not normalize the loudness of SpeechDat corpus, so there are different loudness levels in test and train datasets. Table 1 shows that RASTA LC-RC features give the best performance (phoneme error rate (PER) reduction is from 45.5% to 44.1%). Normalized C_0 gave a significantly poorer result.

Table 1. PER for different front-ends

	Baseline LC-RC	RASTA LC-RC	New C0 LC-RC
PER	45.5	**44.1**	45.9

In the second experiment we evaluated the dependency between speech loudness and recognition accuracy. All files in the test set were normalized to maximum value (at 16 bit it was 32000). After that, the phoneme recognition performance was estimated. The performance estimation was repeated with the test set at -5, -10, -15, -20, -25 dB (relative to maximum loudness).

Table 2 shows that the RASTA-like LC-RC gives the best robustness to speech volume (PER decreases from 52.1 to 49.2 at -25 dB). The RASTA-like filter decreases PER about 1-4% at different loudness.

The results of the normalized C_0 (New C_0) were unexpected. These features gave better performance for loud speech, but for lower loudness levels they demonstrate higher PER than the baseline.

All types of ASR give the best performance at the loudness level of -10 dB. It can be explained by the fact that the average speech loudness of the training set is the same. This fact can be used in training strategy – better ASR performance can be achieved when training and test sets have the same loudness.

Table 2. PER as a function of speech loudness for the proposed front-end

	-25 dB	-20 dB	-15 dB	-10 dB	-5 dB	Max
Baseline LC-RC	52.1	47.3	45.0	**44.8**	47.0	52.8
RASTA LC-RC	49.2	45.6	43.9	**43.8**	45.2	48.8
New C0 LC-RC	54.4	49.2	46.3	**45.7**	47.2	51.7

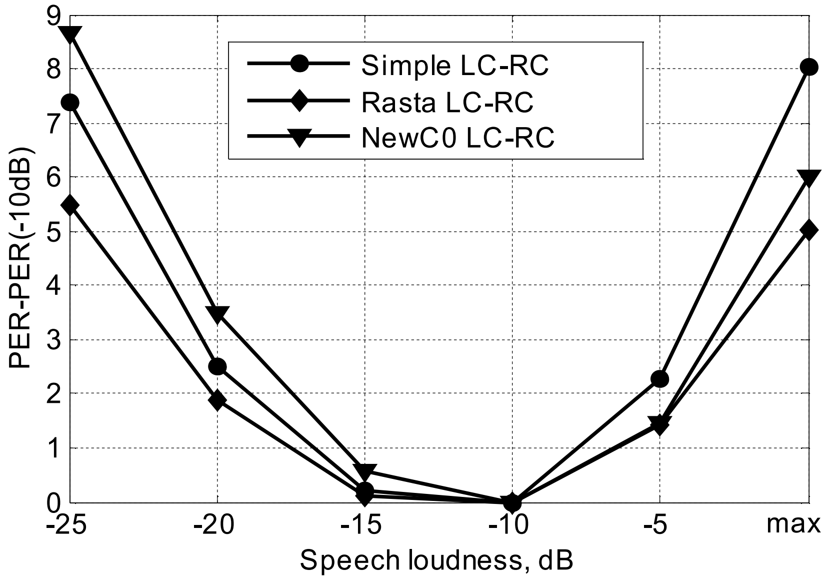

Figure 3. PER as a function of the loudness level.

Figure 3 shows the dependency of PER on speech loudness. For the sake of demonstration PER at -10 dB has been subtracted from all values in Table 2.

7. Conclusions

We have proposed enhancements of the LC-RC front-end and demonstrated that they lead to a significant improvement of ASR robustness to speech loudness. RASTA-like filtering reduces PER by 1% at a normal loudness level and by up to 4% when speech volume is changing. The C_0 normalization technique did not give any improvement. We can explain this by the fact that changes in speech loudness do not have an equal effect on every critical band, so C_0 subtraction must depend on frequency.

It should be noted that the proposed RASTA-like filter only reduces PER and does not provide absolute insensitivity to loudness, so we are going to continue our research in this field.

References

1. H. Hermansky; S. Sharma, "Temporal patterns (traps) in asr of noisy speech", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP), Phoenix, Arizona, USA, Mar. 1999.
2. P. Jain; H. Hermansky, "Beyond a single critical-band in trap based asr", in Proc. Eurospeech, Geneva, Switzerland, Sep. 2003.
3. F. Grezl, "TRAP-based probabilistic features for automatic speech recognition", Ph.D. thesis, Brno University of Technology, 2007.
4. P. Schwarz, "Phoneme recognition based on long temporal context", Ph.D. thesis, Brno

University of Technology, 2008.

5. P. Matejka, P. Schwarz, J. Chernocky, P. Chytil, “Phonotactic language identification high quality phoneme recognition,” in . Eurospeech, 2005, pp. 2237–2240.
6. H. Hermansky, “Auditory modeling in automatic recognition of speech”, in Proceedings of the First European Conference on Signal Analysis and Prediction, pp. 17-21, Prauge, Czech Republic, 1997.
7. H. Hermansky and N. Morgan, “RASTA processing of speech”, IEEE Trans. Speech and Audio, 2:578-589, October 1984.
8. Ivan Tampel, Ekaterina Lysenko «A novel post-filter for speech recognition features» Труды IX Международной научно-технической конференции “ФИЗИКА И РАДИОЭЛЕКТРОНИКА В МЕДИЦИНЕ И ЭКОЛОГИИ” (ФРЭМЭ’2010).