IDIAP RESEARCH REPORT

NON-LINEAR SPECTRAL CONTRAST STRETCHING FOR IN-CAR SPEECH RECOGNITION

Weifeng Li [a] Herve Bourlard [a]

IDIAP–RR 07-53

SEPTEMBER 2007

PUBLISHED IN
Interspeech 2007

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-53

Non-linear Spectral Contrast Stretching for In-car Speech Recognition

Weifeng Li Herve Bourlard

September 2007

published in
Interspeech 2007

Abstract. In this paper, we present a novel feature normalization method in the log-scaled spectral domain for improving the noise robustness of speech recognition front-ends. In the proposed scheme, a non-linear contrast stretching is added to the outputs of log mel-filterbanks (MFB) to imitate the adaptation of the auditory system under adverse conditions. This is followed by a two-dimensional filter to smooth out the processing artifacts. The proposed MFCC front-ends perform remarkably well on CENSREC-2 in-car database with an average relative improvement of 29.3% compared to baseline MFCC system. It is also confirmed that the proposed processing in log MFB domain can be integrated with conventional cepstral post-processing techniques to yield further improvements. The proposed algorithm is simple and requires only a small extra computation load.

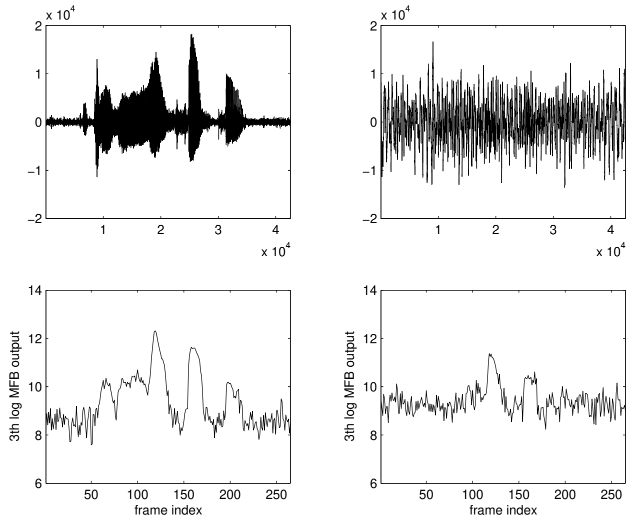

Figure 1: Effect of car noise on log mel-filter bank (MFB) trajectories. Upper left: waveform of close-talking speech; Lower left: log MFB output of the third channel for close-talking speech; Upper right: waveform of distant speech; Lower right: log MFB output of the third channel for distant speech. The speech is "12439" in Japanese.

1 Introduction

The auditory system is particularly responsive to spectral changes and seems particularly well suited to such kind of changes [1]. When some aspect of a stimulus is changed, the changed aspect stands out and thus will be enhanced perceptually. As an example, auditory-nerve fibers show increased rates of firing at the onsets of syllables and during transient events such as stop constant burst. It has been suggested that such adaptation of the auditory system plays an important role in enhancing the *spectral contrast* under adverse conditions [2]. From the formal experiments performed by Summerfield and his colleagues [3], it is shown that the perception of speech-like sounds depends on the *spectral difference* between the current sound and the preceding sound.

Standard mel-frequency cepstral coefficients (MFCC) [4] are extracted from log scaled mel-filterbank (MFB) outputs. Under adverse conditions, background noise generally leads to a reduction in the dynamic ranges of log MFB trajectories. This kind of reduction implies a decay of spectral contrast between speech and background noise. Figure 1 shows the third-channel log MFB trajectories of speech captured by a close-talking (headset) microphone and a distant microphone (attached to the ceiling above the driver's seat [5]) in a car-driving environment. Compared to close-talking speech, the floor level of the log MFB trajectory of distant speech is elevated and the valleys are buried by noise energy. While spectral contrasts between some speech segments and noise are rather evident for close-talking speech, they disappear for distant speech. If the derived MFCC front-ends are fed into an automatic speech recognition (ASR) system, they will produce a mismatch between relatively clean speech (for training) and noisy speech (for testing). Motivated by the adaptation of the auditory system described above, we therefore propose to enhance the spectral contrast of the log mel-filterbank (MFB) outputs. We term this "*spectral contrast stretching*". In order to smooth out the processing artifacts, a two-dimensional smoothing filter is applied. Finally, the corresponding cepstral features are computed. Through spectral contrast stretching and two-dimensional smoothing, the mismatch between the training and test conditions can be reduced, which helps improve the performance of an ASR system.

It is well known that the distortions caused even by additive noise are highly non-linear in the log spectral or cepstral domain. Although linear methods (like the Cepstral Mean Normalization (CMN) or Mean and Variance Normalization (MVN) [6]) are shown to provide significant improvements for

ASR systems, these methods have limitations in non-linear distortion. Histogram equalization (HEQ) [7] provides a non-linear compensation, however, it depends on an explicit assumption of the reference probability distribution and/or the accuracies of the estimated histograms. Our proposed method , which is oriented to the compensation of the mismatch between training and test conditions, can deal with the non-linear effects and does not have the limitations of HEQ techniques. Furthermore, our method can be integrated with conventional cepstral post-processing techniques, such as CMN, MVN, or HEQ, to yield further improvements. Although the present implementation is in the standard log MFB domain, the proposed method can be straightforward applied for J-RASTA [8] or in the root power domain [9].

The organization of this paper is as follows: Section 2 describes the proposed MFCC front-ends. In section 3, the experimental evaluations are presented, and section 4 draws the conclusions.

2 The proposed MFCC front-ends

The proposed mel-frequency cepstral coefficients (MFCC) are calculated according to the scheme shown in Figure 2. Firstly overlapping frames of short-time speech are extracted, pre-emphasized, and filtered by a zero-padded Hamming window. Then the fast Fourier transformation (FFT) is performed. Mel-scaled bandpass filters are applied by weighting the power spectrum FFT coefficients. Then the logarithmic outputs of the filterbanks are computed. The following processing blocks are explained in the next subsections.

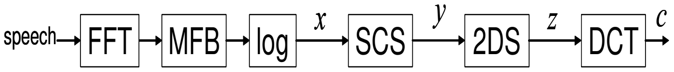

Figure 2: Block diagram of the proposed mel-frequency cepstral coefficients (MFCC). The abbreviations are: FFT - fast Fourier tranformation; MFB - mel-scaled filterbank anaysis; SCS - spectral contrast streching; 2DS - 2-D smoothing; DCT - discrete cosine transformation.

2.1 Non-linear spectral contrast stretching (SCS)

Non-linear spectral contrast stretching (SCS) is performed for each mel-filterbank. Let $x(k,l)$ denote the log mel-filterbank output at the k-th filterbank channel and the l-th frame. Let $x_{\max}(k)$ denote the maximum value of the kth filterbank along the frame index, i.e.,

$$x_{\max}(k) = \max[x(k,1), x(k,2), \ldots, x(k,L)], \tag{1}$$

where L is the number of the frames for an utterance. Let $x_n(k)$ denote the estimated noise in the log mel-filterbank (MFB) domain. More specifically, $x_n(k)$ can be estimated using the first P noise (non-speech) frames, i.e.,

$$x_n(k) = \frac{1}{P}\sum_{l=1}^{P} x(k,l). \tag{2}$$

The non-linear compensation is implemented by

$$y(k,l) = \frac{U(x(k,l) - x_n(k))}{x_{\max}(k) - x_n(k)} \cdot x(k,l), \tag{3}$$

where $U(\cdot)$ is the step function:

$$U(v) = \begin{cases} v, & \text{if } v \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

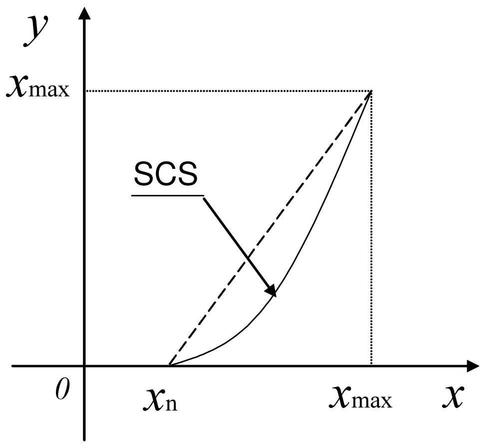

Figure 3: Graphic representation of spectral contrast stretching (SCS) implemented by Eq. (3).

A graphic representation of spectral contrast stretching (SCS) based on Eq. (3) is shown in Figure 3 (assuming $x_n > 0$). When $x(k,l) \leq x_n(k)$, the proposed transformation implicitly includes a noise subtraction in the log MFB domain. When $x(k,l) > x_n(k)$, the curve is monotonically increasing and takes a non-linear quadratic form. The proposed transformation tends to amplify speech changes (likely appear as peaks) and to suppress noise components (likely appear in the valleys). In this way, the spectral contrast beween the speech and the noise can be enhanced. The proposed algorithm also has a function of variance normalization in that the dynamic range is normalized to $[0, x_{\max}(k)]$.

2.2 2-D smoothing (2DS)

By taking into account the time sequence and the log mel-filterbank vector jointly, a two-dimensional (2-D) time-filterbank matrix is obtained. In order to eliminate the processing artifacts due to the previous operations, the 2-D time-filterbank matrix is filtered by using the 2-D smoothing filters which are widely used in image restoration.

Mean filtering [10] is a simple but effective method of smoothing images. Mean filtering simply replaces each pixel value in an image with the mean value of its neighbors, including itself. In our case, the 2-D log mel-filterbank ouput $z(k,l)$, smoothed by using a mean filter, can be obtained as

$$z(k,l) \quad = \quad \frac{1}{MN} \sum_{(m,n) \in R} y(m,n), \tag{5}$$

where R denotes the window used $M \times N$ and $y(m,n)$ denotes the mel-filterbank outputs around the point (k,l), which is obtained in the previous subsection. Given a 3×3 window, the mean filter can be represented as

$$w(k,l) = \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{6}$$

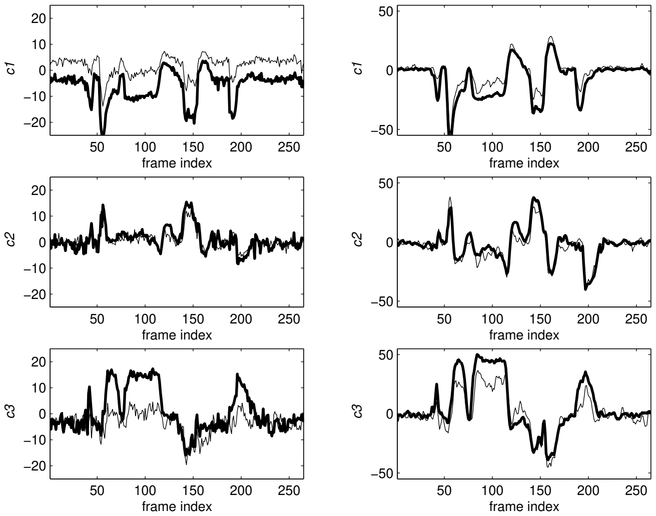

Figure 4: 1-3th MFCC trajectories of the close-talking speech and the distant speech without and with compensation (The speech is the same as Figure 1). The left three small figures depict the original versions with standard MFCCs, and the right three small figures correspond to the ones using the proposed MFCCs. Inside each small figure, bold line is for clean speech and thin line is for distant speech.

2.3 Discrete cosine transformation (DCT)

The proposed Mel-frequency cesptral coefficients (MFCC) are calculated from $z(k, l)$ using the discrete cosine transform (DCT)

$$c(i, l) = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} z(k, l) \cos\left(\frac{\pi i}{K}(k - 0.5)\right) \tag{7}$$

where i and K denote the MFCC index and the number of filterbank channels, respectively.

The 1-3th MFCC trajectories for the original and the compensated are plotted in Figure 4 (The speech is the same as Figure 1). As shown in this figure, original versions yield remarkable mismatches between the close-talking speech and the distant speech in the first and third MFCC trajectories. By using the proposed method, however, the mismatches are reduced and the speech variations become more pronounced. As for the second MFCC trajectories which are almost matched, the proposed method can keep the matched property. Note that the dynamic ranges of the compensated MFCC trajectories become larger because of the contrast stretching in the log MFB domain.

3 Speech Recognition Experiments

3.1 Experimental setup

In order to evaluate the proposed algorithms, the CENSREC-2 database [5] was used to perform speech recognition experiments. This database comprises a task for continuous digit recognition in real car driving environments. In-car speech data is collected in a specially equipped vehicle under 11 environmental conditions. The speech recorded by a distant microphone (attached to the ceiling above the driver's seat) is used for evaluation. There are four evaluation environments for speech recognition depending on whether the recording environments and the microphones used between training and testing data are matched or not.

Table 1: Relative improvements (RI) (as percentages) for different methods. The first line denotes word accuracies (as percentages) for the baseline recogntion system [5].

baseline	81.6	74.5	61.5	48.9	66.4
	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
LSA	0.6	21.9	16.3	22.2	17.4
AFE	17.4	-16.0	1.9	14.5	5.5
CMN	2.0	23.2	0.2	0.1	4.8
MVN	7.1	27.1	23.6	29.8	24.2
HEQ	15.1	23.7	14.3	19.7	18.3
proposed	6.2	34.8	32.9	32.6	29.3
+CMN	12.1	33.1	39.0	48.5	**37.6**
+MVN	10.1	36.9	34.9	44.3	35.3
+HEQ	21.4	37.2	35.4	37.3	34.4

Table 2: Word accuracies (as percentages) achieved by combining the proposed MFCC front-ends with cepstral mean normalization (CMN).

car speed	in-car condition	condition 1	condition 2	condition 3	condition 4
idling	normal	94.6	–	–	–
	fan on	95.0	–	–	–
	audio on	63.6	–	–	–
	window open	96.6	–	–	–
low speed (on city street)	normal	91.6	92.2	87.5	86.2
	fan on	91.8	91.6	85.7	83.3
	audio on	71.8	76.9	71.9	69.2
	window open	80.2	83.4	74.5	70.8
high speed (on expressway)	normal	85.4	84.1	77.5	73.0
	fan on	84.6	82.1	74.7	70.3
	audio on	67.7	69.9	63.4	62.4
overall		82.9	82.9	76.5	73.7

The speech signals are sampled at 16 kHz. In the baseline system, spectral components lower than 250 Hz are filtered out to compensate for the spectrum of engine noise, which is concentrated in the lower frequency region. The duration of the analysis window is 20 milliseconds with a frame shift of 10 milliseconds. 12 mel-frequency cepstral coefficients (MFCC) are drived based on a 24-channel MFB analysis. Finally, a vector size of 39 parameters, including 12 MFCC, log energy, and their deltas and the accelerations, is used for HMM modeling. Further details of the corpus and the baseline speech recognition system can be found in [5].

3.2 Speech recognition results

Table 1 shows the recognition results of different methods in terms of relative improvement (RI), which is calculated as

$$RI(\%) = \frac{NewScore - Baseline}{100 - Baseline} \times 100\ (\%) \tag{8}$$

where $NewScore$ and $Baseline$ are word accuracies (as percentages) for each testing method and reference algorithm, respectively. The upper part of this table depicts the relative improvement by using conventional methods: "LSA" denotes the conventional MFCCs extracted from the speech enhanced using the minimum mean square error (MMSE) on log-spectral amplitude [11]; “AFE” denotes the ETSI advanced front-end [12]; “CMN”, “MVN”, “HEQ” denote the cepstral post-processing methods based on cepstral mean normalization, mean and variance normalization, and histogram equalization, respectively. The lower part of this table corresponds to the ones using the proposed MFCC front-ends and with the combinations of CMN, MVN, and HEQ techniques. CMN, MVN, and HEQ are performed for the 12 static cepstral coefficients, leaving the delta and acceleration coefficients unchanged. In the proposed MFCC front-ends, the 2nd $\sim$ 24th log MFB outputs are smoothed using the mean filter (the first log MFB and log energy outputs are filtered as zero), then the cepstral coefficients are derived using Eq. (7) and the delta and acceleration coefficients are calculated.

As shown in Table 1, the speech enhancement method LSA is effective for all the four evaluation conditions for its noise reduction effects, however for speech recognition, the algorithm introduces much computation cost. The ETSI advanced front-end (AFE) is not very effective except for condition 1 where both the recording environments and the microphones are matched. Using all the three conventional normalization methods (CMN, MVN, and HEQ) in cepstral domain is helpful for improving the in-car speech recognition performance, and among them MVN provides the largest gain (a relative improvement of 24.2% in average). The CMN yields only a marginal gain for the last two conditions where a close-talking headset microphone is used for training and a distant microphone for testing.

Compared with the five conventional methods, the proposed MFCC front-ends perform better in terms of average relative improvement and are very effective for the last three conditions where the conditions between training and test data are highly mismatched. Moreover, it can be found from this table that the combinations of conventional normalization methods for post-processing in cepstral domain yield higher recognition accuracies. The best recognition performance (a relative improvement of 37.6% in average) is achieved by combining the proposed MFCC front-ends with cepstral mean normalization (CMN). The degree of gain provided by MVN , and HEQ is not necessarily consistent in comparing systems with and without the proposed MFCC front-ends. The detailed experimental results of the proposed MFCC front-ends combined with CMN is given in Table 2 as well.

4 Conclusions

In this paper, we have proposed non-linear spectral contrast stretching (SCS) based MFCC front-ends for the improvement of speech recognition. It is evaluated on the CENSREC-2 continuous digit recognition task in real in-car environments. The proposed MFCC front-ends can have an average relative improvement of 29.3% compared to the baseline MFCC system. Further improvements are achieved when the proposed method is integrated with conventional cepstral post-processing techniques. Future directions lies in the investigation of the performance of the proposed method using other corpus and in more elaborate noise estimation with the combination of other speech enhancement methods to further improve the recognition accuracy.

5 Acknowledgements

This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811) and the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2.

References

[1] S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. Fay, Speech Processing in the Auditory System, Springer-Verlag, New York, 2004.

[2] W. J. Hardcastle and J. Laver, The Handbook of Phonetic Sciences, Blackwell Publisher, 1999.

[3] A. Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," Journal of the Acoustical Society of America, Vol.81, No.3, pp.700-708. 1987.

[4] S. B. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoustics, Speech, Signal Processing, Vol.ASSP-28, No.4, pp.357-366, 1980.

[5] S. Nakamura, M. Fujimoto, and K. Takeda, "CENSREC2: Corpus and Evaluation Environments for In Car Continuous Digit Speech Recognition", Proc. ICSLP (Interspeech 2006), pp.2330-2333, 2006.

[6] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition", Speech Communication, Vol.25, pp.133-147, 1998.

[7] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, A.J. Rubio, "Histogram equalization of speech representation for robust speech recognition", IEEE Trans. Speech and Audio Processing, Vol.13, No.3, pp.355-266, 2005.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech and Audio Processing, Vol.2, No.4, pp.578-589, 1994.

[9] P. Alexandre and P. Lockwood, "Root cepstral analysis: a unified view-application to speech processing in car noise environments", Speech Communication, Vol.12, pp.277-288, 1993.

[10] A. K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, 1989.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-33, No.2, pp.443-445, 1985.

[12] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," ETSI ES 202 050 v1.1.1, 2002.