**Bagher BABA ALI[1], Waldemar WÓJCIK[2], Orken MAMYRBAYEV[3], Mussa TURDALYULY[3], Nurbapa MEKEBAYEV[3]**

University of Tehran (1), Iran
Lublin University of Technology (2),
Institute of Information and Computational Technologies (3), Kazakhstan

doi:10.15199/48.2018.06.17

# Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction

***Abstract***. *Spectral compression is an effective robust feature extraction technique to reduce the mismatch between training and testing data in feature domain. In this paper we propose a new MFCC feature extraction method with non-uniform spectral compression for speech recognition in noisy environments. In this method, the energies of the outputs of the mel-scaled band pass filters are compressed by different root values adjusted based on information from the back-end of speech recognition system. Using this new scheme of speech recognizer based non-uniform spectral compression (SRNSC) for mel-scaled filter-bank-based cepstral coefficients, substantial improvement is found for recognition in presence of different additive noises with different SNR values on TIMIT database, as compared to the standard MFCC and features derived with cubic root spectral compression.*

***Streszczenie.*** *Kompresja spektralna jest efektywną i niezawodną techniką wyodrębniania cech w celu zmniejszenia niedopasowania między danymi uczącymi i testowymi w domenie cech. W tym artykule proponujemy nową metodę wyodrębniania cech MFCC z niejednorodną kompresją spektralną do rozpoznawania mowy w hałaśliwym otoczeniu. W opisywanej metodzie, energie wyjść pasmowych filtrów skali melowej są kompresowane przez różne wartości bazowe wyznaczone na podstawie informacji z back-endu systemu rozpoznawania mowy. Stosując ten nowy schemat niejednorodnej kompresji spektralnej (SRNSC) opartej na rozpoznawaniu mowy dla współczynników cepstralnych opartych na banku filtrów o skali melowej, stwierdzono znaczną poprawę rozpoznawania w obecności różnych szumów addytywnych o różnych wartościach SNR z bazy danych TIMIT, w porównaniu do standardowego MFCC i cech wyznaczonych za pomocą pierwiastkowej kompresji spektralnej. (**Niejednorodna kompresja spektralna do odpornej ekstrakcji cech MFCC**).*

**Keywords:** Robust speech recognition, Spectral compression, Speech recognizer-based.
**Słowa kluczowe:** odporne rozpoznawanie mowy, kompresja spektralna.

## Introduction

The performance of speech recognition degrades dramatically in the presence of noise, due to spectral mismatch between the training and testing data. Therefore, robust speech recognition in noisy environment is still a challenging problem. To solve this problem, many compensation techniques have been proposed by researchers. In general, a compensation technique can be applied in the signal, feature or model space [1].

This paper focuses on the compensation in feature domain. Spectral compression is an effective robust feature extraction technique to reduce the mismatch between training and testing data in feature domain. In conventional Mel-frequency cepstral coefficients (MFCC) feature extraction, a logarithm function is applied to Mel filter bank energies in order to reduce their dynamic range. Root cepstral analysis [2] replace log function with a constant root function and yields RCC coefficients. RCC coefficients have shown better robustness against the noise. In RCC method compressed speech spectrum is computed as shown in (1):

$$P_C(m) = P(m)^{\gamma} \quad , \quad 0 \le \gamma \le 1 \tag{1}$$

where $P_C(m)$ is the compressed spectrum, $P(m)$ is the original spectrum, $\gamma$ is the compression factor and $m$ is the filter bank index.

In (1), the compression factor is fixed for all the frequency bands under the assumption that the noise contamination is same throughout all frequency bands, although real world noise is mostly colored and does not affect the speech signal uniformly over the entire spectrum. Therefore, the compression factor should be adjusted for each band. Also, from the psychoacoustic point of view, using constant compression root for all frequencies is sub-optimal [3]. Therefore, relation (1) is extended as follows:

$$P_C(m) = P(m)^{\gamma(m)} \quad , \quad 0 \le \gamma(m) \le 1 \tag{2}$$

where the compression factor is dependent on the frequency band and named non-uniform spectral compression. Even if the noise is stationary, this parameter should be calibrated at the beginning. In conventional methods [3,4,5], the compression factor is adjusted for each band according to the SNR instead of recognition results that seem more conscionable. In SNR-based approaches there is no feedback from recognition stage to the compensation stage and they explicitly need to SNR estimation block. Thus, its performance depends on SNR estimation accuracy. However, the speech recognition is a classification problem and this seems reasonable that any adjustment in parameters of compensation techniques is resulted in improvement in recognition performance [6,7]. Compensation method improves speech recognition accuracy, only when it generates the sequence of feature vectors which maximize the likelihood of the correct transcription with respect to other hypothesis. Therefore, it seems logical that each parameters calibration of compensating techniques in front-end stage of the speech recognition systems be according to the recognition criteria instead of waveform level criteria such as signal to noise ration.

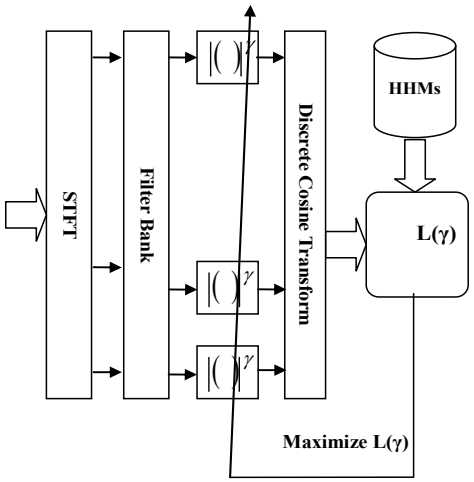


Fig.1. An example of the figure inserted into the text

In this paper a novel framework for applying non-uniform spectral compression in front-end of the speech recognition systems is proposed. We show that by incorporating the speech recognition system into compression factor adjusting process, the recognition rate is further improved. For implementing this scheme, we use an utterance the transcription of which is given and formulate the relation between compression factor and likelihood of the correct model. The proposed method has two phases, adaptation and decoding. In adaptation phase, compression factor is adjusted based on maximizing acoustic likelihood of the correct transcription and in decoding phase this optimized compression factor is applied for all incoming speech. Fig. 1 shows our proposed method for removing noise effects from MFCC features in speech recognizer-based framework. The reminder of this paper is organized as follows. In the next section we derive the framework for speech recognizer based spectral compression. Algorithm of proposed framework is described in the third section. Extensive experiments to verify the effectiveness of presented framework are presented in the following section and finally we present summary of our work in scjematically depicted in Fig. 1.

**Speech Recognizer Based Non-Uniform Spectral Compression (SRNSC).**

Conventional spectral compression techniques use the waveform-level criteria such as signal to noise ratio to calibrate its parameter. According to reasons are mentioned in the introduction, using recognition error rate criteria instead of waveform level criteria for adjusting spectral over-subtraction parameters seems more promising. One logical way for achieving to this goal is to select spectral compression factor so as to maximize acoustic likelihood of the correct hypothesis. This will increase the distance between the acoustic likelihood of the correct hypothesis and other competing hypothesis, so the probability that utterance be correctly recognized will be increased. Hence, in following the relation between spectral compression vector in pre-processing stage with acoustic likelihood of the correct hypothesis in decoding stage is formulated. These formulas depend on feature extraction algorithm and acoustic unit model. In this work, MFCC algorithm and hidden Markov model with mixture of Gaussian in each state are used for feature extraction and modeling of the acoustic unit respectively.

Speech recognition systems based on the statistical model find the acoustic unit sequence most likely to generate observed feature vectors $Z = \{\mathbf{z}_1,...,\mathbf{z}_t\}$ extracted from the improved speech signal. These observed features are a function of both incoming speech signal and also spectral compression vector. Speech recognizer gets the most likely hypothesis based on the optimal Bayes classification formula:

$$\hat{w} = \arg\max_{w} P(Z(\boldsymbol{\gamma})|w)P(w) \tag{3}$$

where the observed feature vectors is a function of spectral compression vector $\boldsymbol{\gamma}$. In this formula $P(Z(\boldsymbol{\gamma}) \mid w)$ and $P(w)$ are the corresponding acoustic and language score, respectively. Our goal is to find vector $\boldsymbol{\gamma}$ achieving the best recognition performance. Similar to either speaker or environmental adaptation methods, for adjusting $\boldsymbol{\gamma}$, we need some adaptation data with a known transcription. We assume that the correct transcription of the utterance $w_C$ is known. So the value of $P(w_C)$ can be ignored because this value is constant regardless of the value of $\boldsymbol{\gamma}$. We can then maximize equation (3) with respect to $\boldsymbol{\gamma}$ as:

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\gamma}(P(Z(\boldsymbol{\gamma})|w_C)) \tag{4}$$

In an HMM based speech recognition the acoustic likelihood is estimated by single most likely state sequence. If $S_C$ represents all state sequences in the combinational HMM and $s$ represents single most likely state sequence, then the Maximum likelihood estimation of $\boldsymbol{\gamma}$ is written as:

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}, s\in S_C}\left\{\sum_i \log(P(z_i(\boldsymbol{\gamma})|s_i)) + \sum_i \log(P(s_i|s_{i-1}, w_C))\right\} \tag{5}$$

Regarding to (5) for getting $\hat{\boldsymbol{\gamma}}$ acoustic likelihood of the correct transcription should be jointly maximize with respect to the state sequence and $\boldsymbol{\gamma}$ parameters. This joint optimization should be performed in an iterative manner.

Noisy speech is passed to the spectral subtraction filter and feature vector $Z(\boldsymbol{\gamma})$ is extracted given the known $\boldsymbol{\gamma}$. Then optimal state sequence $s = \{s_1,...,s_t\}$ is computed using (5), given the correct transcription $w_C$. State sequence $\hat{s}$ simply can be computed using Viterbi algorithm. Given the known state sequence $\hat{s}$ we want to find $\hat{\boldsymbol{\gamma}}$ as:

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\gamma}\left\{\sum_i \log(P(z_i(\boldsymbol{\gamma})|\hat{s}_i))\right\} \tag{6}$$

There is not a closed-form solution for computing optimal $\boldsymbol{\gamma}$ for a given state sequence, so we use non-linear optimization.

**Algorithm**

In this section we present a novel approach for adjusting spectral compression vector so as to maximize acoustic likelihood of the correct transcription. Here we should answer a question. If the correct transcription was known a priori, there would be no need for recognition. For answering this question we should mention that the correct transcription is only needed in the adaptation phase and in the decoding phase these parameters are fixed. At first user says an utterance with a priori known transcription then the utterance is passed through the spectral compression block fixed with initial parameters. After that most likely state sequence is generated using Viterbi algorithm. Then the optimum spectral compression vector is produced given the state sequence. Recognition is performed on validation set with using this optimized filter if the desired error rate is satisfied, the algorithm is finished otherwise the new state sequence is estimated. In an iterative manner, the spectral compression vector which maximizes the total log likelihood of the utterance with a known transcription is found. Feature vector at first is extracted from the improved speech signal and then the log likelihood is computed given the known state sequence. If the likelihood does not converge, the gradient of the spectral compression vector is computed, and the spectral compression vector is updated. Spectral compression is performed with using this updated vector and the new feature vectors are extracted. This process repeats until the convergence condition is satisfied.

In proposed algorithm like speaker and environment adaptation techniques, adaptation of spectral compression vector can be implemented either in a separate off-line session or by embedding an incremental on-line step to the normal recognition mode of system. In off-line adaptation as

explained in above, the user is aware of adaptation, typically by performing a special adaptation session, while in on-line adaptation the user may not even know that adaptation is carried out in Fig. 2 and Fig. 3.

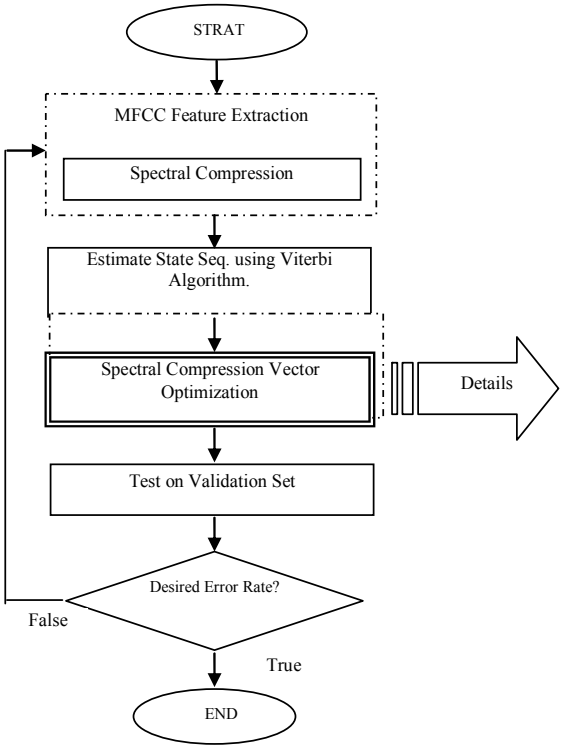


Fig.2. Flowchart of the proposed algorithm

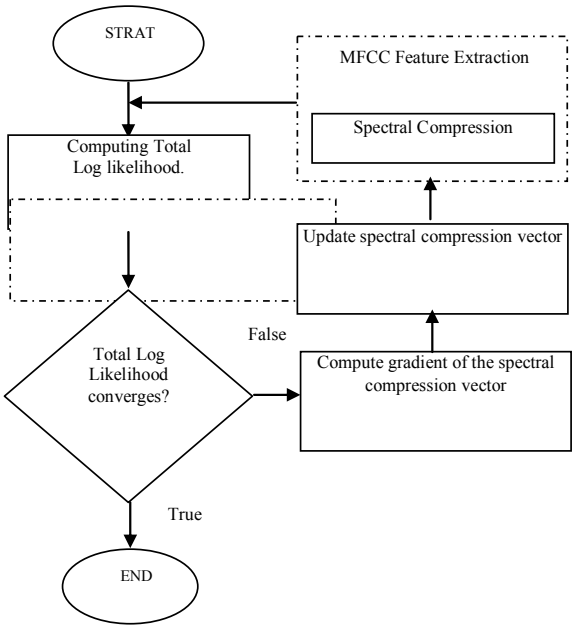


Fig.3. Flowchart of spectral compression vector optimization

**Experiments**

This section presents the experiments evaluations of the proposed algorithms. Throughout this paper, phoneme error rate is used to evaluate the performance of the proposed algorithm. The error rate is computed as follows:

$$\text{Error}(\%) = \frac{\text{Sub} + \text{Del} + \text{Ins}}{\text{Total Number of Phonemes}} \tag{7}$$

where “Sub” is the number of substitutions, “Del” is the number of deletions, and “Ins” is the number insertions.

Database Definition. In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the TIMIT database. The test sets is obtained by the artificially adding of five types of noises (Alarm, Brown, Multi-Talk, Volvo and White noise) from the NOISEX-92 database [6] to a subset of the test set of TIMIT database. For each noise type, the testing data were contaminated at several SNRs from 0dB to 20dB at the interval of 5dB to produce various noisy test sets. Sentences were corrupted by adding noise scaled on a sentence-by-sentence basis to an average power value computed to produce the required SNR.

Baseline Speech Recognizer. Speech recognition experiments conducted using NEVISA [8], a large vocabulary, speaker-independent, continuous HMM-based speech recognition system has been developed at speech processing lab of computer engineering department of Sharif University of technology. Experiments have been done in the phoneme recognition operational modes of NEVISA system. The reason for performing phoneme recognition instead of word recognition is that in the former case, the recognition performance lies primarily on the acoustic model. For word recognition, the performance becomes sensitive to various factors for example the language model type.

Each phoneme was modeled by six state continuous density left-to-right HMMs. In addition, silence was explicitly modeled by a HMM. The observation densities were mixtures of eight multivariate Gaussian distributions with diagonal covariance matrices. Forward and skip transitions between the states and self-loop transitions were allowed. Thirty-six dimensional feature vectors were used: $C[1]$ to $C[12]$ derived from a mel-spaced filter-bank of 25 filters, and their first and second derivatives to make up vectors of 36 coefficients per speech frame.

**Recognition Results**

In all of our experiments, one sentence of test set is used for optimization phase. After the spectral compression vector was optimized, speech recognition is performed on other sentences of test set using this optimized vector. Table 1 shows phoneme error rate in each test condition. For evaluating our algorithm, its results are compared with MFCC and RCC features which derived with log and cubic root spectral compression.

Table 1. Phoneme recognition accuracy (%) on TIMIT database

| Noise Type | Method | 0dB | 5dB | 10dB | 15dB | 20dB |
|---|---|---|---|---|---|---|
| Alarm | MFCC | 19.46 | 28.95 | 37.06 | 46.07 | 52.91 |
| | RCC | 18.16 | 27.47 | 38.21 | 47.06 | 53.33 |
| | SRNSC | 20.42 | 30.77 | 39.48 | 48.15 | 54.37 |
| Brown | MFCC | 40.6 | 48.53 | 56.46 | 61.92 | 63.96 |
| | RCC | 40.52 | 48.89 | 57.29 | 63.16 | 64.77 |
| | SRNSC | 42.12 | 50.31 | 57.56 | 63.98 | 65.05 |
| Multi-Talk | MFCC | 13.99 | 23.6 | 34.35 | 43.78 | 51.89 |
| | RCC | 13.04 | 23.8 | 35.05 | 43.99 | 52.45 |
| | SRNSC | 14.98 | 24.17 | 35.44 | 45.17 | 53.59 |
| Volvo | MFCC | 44.44 | 48.59 | 53.39 | 57.36 | 61.02 |
| | RCC | 43.94 | 48.65 | 53.53 | 58.93 | 62.19 |
| | SRNSC | 46.25 | 51.51 | 55.53 | 59.38 | 63.04 |
| White | MFCC | 3.72 | 8.47 | 15.68 | 23.96 | 31.77 |
| | RCC | 2.11 | 6.61 | 14.90 | 24.72 | 31.98 |
| | SRNSC | 3.86 | 9.91 | 16.38 | 25.31 | 33.09 |

From the experimental results, we observed the following facts. With regards to the difference between noise type and SNR, the result shows that the proposed method was capable to improve recognition performance compared to classical methods. In some cases, RCC method achieves lower performance than the baseline (MFCC). This is due to some frequency components under-compressing or over-compressing caused by not adjusting

the spectral compression vector and destroy the discriminability in pattern recognition. The mismatch reduces the effectiveness of the clean trained acoustical models and causes the recognition accuracy to fall. Table 1 Phoneme recognition accuracy (%) on TIMIT database.

**Conclusions**

We proposed a speech recognizer based non-uniform spectral compression instead of conventional methods. Experimental results on the TIMIT speech database have revealed the effectiveness of it in presence of different additive noises with different SNR values. As our future work, we plan to combine our proposed algorithm with other robustness methods and test it in real conditions.

*Acknowledgments. Author would like to give special thanks to the members of R&D department of ASR-Gooyesh Pardaz Company for their scientific supports and providing a framework to do his experiments.*

***Authors**: Bagher BabaAli, School of Statistics, Mathematics and Computer Science, College of Science, University of Tehran, 16th Azar St., Enghelab Sq., Tehran, Iran, E-mail: babaali@ut.ac.ir; Waldemar Wójcik, Lublin University of Technology, Institute of Electronics and Information Technology, Nadbystrzycka 38A, 20-618 Lublin, Poland, E-mail: waldemar.wojcik@pollub.pl; Orken Mamyrbayev, Institute of Information and Computational Technologies, Pushkin 125, 050010 Almaty, Kazakhstan, E-mail: morkenj@mail.ru; Mussa Turdalyuly, Institute of Information and Computational Technologies, Pushkin 125, 050010 Almaty, Kazakhstan, E-mail: mkt_001@mail.ru; Nurbapa Mekebayev, Institute of Information and Computational Technologies, Pushkin 125, 050010 Almaty, Kazakhstan, E-mail: nurbapa@mail.ru.*

REFERENCES

[1] Acero A., Stern R. M., Robust speech recognition by normalization of the acoustic space, *in IEEE International Conference on Acoustics, Speech, and Signal Processing.*, (1991), 893-896

[2] Alexandre P., Lockwood P., Root cepstral analysis: a unified view: application to speech processing in car noise environments, *Speech Communication*, 12 (1993), 277-288

[3] Chu K. K., Leung S. H., SNR-dependent non-uniform spectral compression for noisy speech recognition, *in IEEE International Conference on Acoustics, Speech, and Signal Processing*, (2004)

[4] Nasersharif B., Akbari A., SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features, *Pattern Recognition Letters*, 28 (2007)

[5] Nafalski A., Wibawa A.P., Machine translation with Javanese speech levels' classification, IAPGOS, 6 (2016), No. 1, 21-25

[6] Varga A., The Noise-92 Study on the Effect of Additive Noise on Automatic Speech Recognition, *DRA Speech Research Unit, St. Andrew's Rd., Malvern, Worcestershire, WR14 3PS UK*, (1992).

[7] Kamińska D., Pelikant A., Spontaneous emotion recognition from speech signal using multimodal classification, IAPGOS, 2 (2012), No. 3, 36-39

[8] Sameti H., Veisi H., Bahrani M., Babaali B., Hosseinzadeh K., NEVISA, A Persian Continuous Speech Recognition System, *in 13th International CSI Computer Conference, Kish Island, Persian Gulf, Iran*, (2008).