Noise-Robust Speech Feature Processing with Empirical Mode Decomposition

Kuo-Hau Wu* and Chia-Ping Chen

[1]Department of Computer Science and Engineering National Sun Yat-Sen University 70 Lien-Hai Road, Kaohsiung, Taiwan 800

Email: Kuo-Hau Wu - wkh97m@cse.nsysu.edu.tw; Chia-Ping Chen - cpchen@cse.nsysu.edu.tw;

*Corresponding author

Abstract

In this paper, a novel technique based on the empirical mode decomposition methodology for processing speech features is proposed and investigated. The empirical mode decomposition generalizes the Fourier analysis. It decomposes a signal as the sum of the intrinsic mode functions. In this work, we implement an iterative algorithm to find the intrinsic mode functions for any given signals. We design a novel speech feature post-processing method based on the extracted intrinsic mode functions to achieve noise-robustness for automatic speech recognition. Evaluation results on the noisy-digit Aurora 2.0 database show that our method leads to significant performance improvement. The relative improvement over the baseline features increases from 24.0% to 41.6% when the proposed post-processing method is applied on the mean-variance normalized speech features. The proposed method also improves over the state-of-the-art performance achieved by the Aurora advanced front-end. The relative improvement advances from 67.2% to 68.2% when the proposed method is inserted in the overall system.

1 Introduction

State-of-the-art automatic speech recognition (ASR) systems can achieve satisfactory performance under well-matched conditions. However, when there is a mismatch between the train data and the test data, the performance often degrades quite badly. The versatility of everyday environments requires ASR systems to function well in a wide range of unseen noisy conditions. Therefore, noise-robust speech processing technology

for recognition is an important research topic, both practically and technically.

Many techniques for noise robustness have been proposed and put to tests. The speech enhancement methods, such as the well-known spectral subtraction [1] and Wiener filters [2], introduce pre-processing steps to remove the noise part or estimate the clean part given the noisy speech signal. The auditory front-end approaches incorporate knowledge of human auditory systems acquired from psychoacoustic experiments, such as the critical bands and the spectral or temporal masking [3, 4], in the process of speech feature extraction. The noise-robust feature post-processing techniques, such as the cepstral mean subtraction (CMS) [5], the cepstral variance normalization (CVN) [6], and the histogram equalization (HEQ) [7], aim to convert the raw speech features to a form that is less vulnerable to the corruption of adverse environments.

In this paper, we study a feature post-processing technique for noise-robust ASR based on the empirical mode decomposition (EMD) [8]. By EMD, a feature sequence (as a function of time) is decomposed by the intrinsic mode functions (IMFs). The low-order IMFs contain high-frequency components and they are removed based on a threshold estimated from the train data in our proposed method.

The organization of this paper is as follows. In Section 2, we introduce the formulation of the empirical mode decomposition and show that it is a generalization of the Fourier analysis. In Section 3, we design an iterative algorithm to extract the intrinsic mode functions for EMD. In Section 4, we describe the proposed EMD-based feature post-processing method and give a few illustrative examples. The experimental results are presented in Section 5, and the concluding remarks are summarized in Section 6.

2 Empirical Mode Decomposition

The empirical mode decomposition generalizes the Fourier series. The sinusoidal basis functions used in the Fourier analysis is generalized to the data-dependent intrinsic mode functions. Compared to a sinusoidal function, an intrinsic mode function satisfies the *generalized alternating property* and the *generalized zero-mean property*, and relaxes the constant amplitude and frequency to become time-varying functions.

2.1 The Fourier Series

A signal, say $X(t)$, of finite duration, say T, can be represented by the Fourier series, which is a weighted sum of the complex exponential functions with frequencies $\omega_k = 2\pi k/T$. That is, we can write

$$\begin{aligned} X(t) &= \sum_{k=-\infty}^{\infty} r_k e^{j\omega_k t} \\ &= r_0 + \sum_{k=1}^{\infty} (r_k + r_{-k}) \cos \omega_k t + j \sum_{k=1}^{\infty} (r_k - r_{-k}) \sin \omega_k t. \end{aligned} \tag{1}$$

Defining

$$p_k = r_k + r_{-k}, \quad q_k = j(r_k - r_{-k}), \quad k = 1, 2, \ldots, \tag{2}$$

we can re-write (1) as

$$X(t) = r_0 + \sum_{k=1}^{\infty} p_k \cos \omega_k t + \sum_{k=1}^{\infty} q_k \sin \omega_k t. \tag{3}$$

If $X(t)$ is real, p_k, q_k in (2) are real. (3) can be seen as a decomposition of $X(t)$ in the vector space spanned by the following basis set

$$\mathcal{B} = \{1\} \cup \{\cos \omega_k t,\ k = 1, 2, \ldots\} \cup \{\sin \omega_k t,\ k = 1, 2, \ldots\}. \tag{4}$$

The following properties of about the basis functions of the Fourier series are quite critical in the generalization to EMD.

- (*alternating property*) The basis functions have *alternating* stationary points and zeros, meaning that between two adjacent stationary points there is exactly one zero, and vice versa.
- (*zero-mean property*) The maxima and minima of the basis functions are opposite in sign, and the average of the maxima and the minima is 0.

2.2 Empirical Mode Decomposition

In the empirical mode decomposition, a real-valued signal $X(t)$ is decomposed by

$$X(t) \approx \sum_k c_k(t). \tag{5}$$

The $c_k(t)$s are called the intrinsic mode functions. As generalization of the sinusoidal functions, an IMF is required to satisfy the following generalized properties.

- (*generalized alternating property*) The number of the extrema (maxima and minima) and the number of zeros may not differ by more than 1.
- (*generalized zero-mean property*) The average of the upper envelope (a smooth curve through the maxima) and the lower envelope (a smooth curve through the minima) is zero.

The amplitude and frequency of an intrinsic mode function are defined as follows. Given a real-valued function $c_k(t)$, let $d_k(t)$ be the Hilbert transform of $c_k(t)$. A complex function $f_k(t)$ is formed by

$$f_k(t) = c_k(t) + j d_k(t) = \alpha_k(t) e^{j(\int \nu_k(t) dt)}. \tag{6}$$

In (6), we identify $\alpha_k(t)$ and $\nu_k(t)$ as the amplitude function and the frequency function of $f_k(t)$. Note that (6) is also a generalization of the Fourier analysis since $\sin \omega_k t$ is the Hilbert transform of $\cos \omega_k t$. Yet, while the sinusoidal functions have constant amplitudes and frequencies, the intrinsic mode functions have time-varying amplitudes and frequencies.

3 Intrinsic Mode Functions

The core problem for empirical mode decomposition is to find the intrinsic mode functions of given signals. In the following subsections, we state the algorithm that we design for EMD and highlight the properties of IMFs with an illuminating instance.

3.1 Algorithm

The following variables are used in the algorithm to extract the intrinsic mode functions of a given function for the empirical mode decomposition.

- $X(t)$: input signal to be analyzed by EMD;
- $c_k(t)$: extracted IMF;
- $r(t)$: remainder function;
- $u(t)$: upper envelope function;
- $l(t)$: lower envelope function;
- $h(t)$: hypothetical function for $c_k(t)$.

The pseudocode and block diagram of the extraction algorithm are given in Figure 7.1 and Figure 7.2. Note that

- The algorithm is iterative. There is an outer loop to control the number of IMFs and there is an inner loop to find the next IMF given the current remainder function.
- The spline interpolation is used to find the envelopes.
- To guard against slow convergence, we enforce a criteria to terminate the iteration if the difference between the old and new candidates of $h(t)$ is below a threshold.

3.2 Illustrative Examples

In the extraction of the intrinsic mode functions, the remainder function $r(t)$ is recursively replaced by the hypothetical function $h(t)$,

$$r(t) \;\leftarrow\; h(t) = r(t) - \frac{1}{2}\left(u(t) + l(t)\right). \tag{7}$$

The envelopes $u(t)$ and $l(t)$ are smoother than $r(t)$ as each envelope is the spline interpolation of a proper subset of points of $r(t)$. Being the remainder after the subtraction of the envelope mean, $h(t)$ approximates the time-varying local high-frequency part of $r(t)$. Whenever $h(t)$ is a valid IMF, it is set to $c_k(t)$ and subtracted, so the remaining part of signal is smoother. Thus, we expect the IMFs to be progressively smooth as k increases.

For an illustrative example, the IMFs extracted from the log-energy sequences of an utterance in the Aurora 2.0 database with a signal-to-noise ratios (SNR) of 0 dB are shown in Figure 7.3. One can see clearly that the degree of oscillation decreases as k increases, as is predicted by our analysis.

4 EMD-Based Feature Post-Processing

The goal of speech feature post-processing is to reduce the mismatch between the clean speech and the noisy speech. In order to achieve this goal, we first look at the patterns introduced by the presence of noises of varying levels, then we propose a method to counter such patterns.

The patterns created by the noises of several SNRs can be observed on the log-energy sequences of an underlying clean utterance in the Aurora 2.0 database, as shown at the top of Figure 7.4. We can see that the oscillation of the speech feature sequence increases with the noise level. That is, the spurious spikes in the sequence basically stems from the noise signal, rather than from the speech signal.

Since the spikes introduced by the noise are manifest in the low-order IMFs, we propose to subtract these IMFs to alleviate the mismatch. That is, for a noisy speech signal $x(t)$ with EMD

$$x(t) = \sum_{k=1}^{K} c_k(t) + r(t), \tag{8}$$

we simply subtract a small number, say N, of IMFs from $x(t)$, i.e.,

$$\hat{x}(t) = x(t) - \sum_{n=1}^{N} c_n(t). \tag{9}$$

At the bottom of Figure 7.4, the EMD post-processed sequences of the same instances are shown. Comparing to the original sequences at the top, we can see that the mismatch between clean and noisy speech is significantly reduced.

5 Experiments

The proposed EMD-based approach to noise-robustness is evaluated on the Aurora 2.0 database [9]. After the baseline results are reproduced, we first apply the commonly used per-utterance mean-variance normalized (MVN) on the speech features to boost the performance, then we apply the proposed EMD-based post-processing to achieve further improvement. Seeing the significant performance gain over the baseline, we apply the proposed method to the Aurora advanced front-end (AFE) speech features [10] to see if further improvement can be achieved on speech features that are already very noise-robust to begin with.

5.1 Aurora Database

The Aurora 2.0 noisy digit database is widely used for the evaluation of noise-robust frontends [9]. 8 additive noises and 2 convolution noises are artificially added and/or convolved to the clean speech data with signal-to-noise ratio (SNR) levels ranging from 20 dB to -5 dB. The multi-train recognizer is trained by a data set (called the multi-train set) consisting of clean and multi-condition noisy speech samples. The clean-train recognizer is trained by a data set (called the clean-train set) consisting of clean speech samples only. Test data in Set A is matched to the multi-condition train data, test data in Set B is not matched to the multi-condition train data, and test data in Set C is further mismatched due to convolution. Note that the proportion of the data amounts of Set A, Set B, and Set C is $2:2:1$. Therefore, the reported averaged word accuracy rate is computed by

$$\mathbf{Avg} = \frac{\mathbf{Set\ A} * 2 + \mathbf{Set\ B} * 2 + \mathbf{Set\ C} * 1}{5}. \tag{10}$$

5.2 Frontend and Backend

The baseline speech feature vector consists of the static features of 13 mel-frequency cepstral coefficients (MFCC) $c_0, c_1, \ldots, c_{12}$. During training and testing, the dynamic features of velocity (delta) and acceleration (delta-delta) are also derived, resulting in a 39-dimension vector per frame.

The backend recognizer uses the standard setting of the Aurora evaluation [9]. That is, we have 16-state wholeword models for the digits, a 3-state silence model, and a 1-state short-pause model. The short-pause state is tied to the middle state of the silence model. The state-emitting probability density is a 3-component Gaussian mixture for a word state, and a 6-component Gaussian mixture for a silence/short-pause state.

5.3 Results

The first set of experiments is a pilot designed to determine which speech feature sequence is to be applied the EMD-based post-processing. Each noisy speech feature sequence is replaced by the corresponding clean speech feature sequence for every noisy utterance of the test data sets, in order to evaluate the impact of data mis-matchedness. Table 8.1 summarizes the results of Aurora 2.0 clean-train tasks with different test SNRs. Replacing the noisy log-energy feature sequences leads to significant improvement in all SNRs. The relative improvement achieves as high as 72.5% in the 10-dB SNR tasks! On average, the relative improvement is 64.2% over 0-20 dB SNR test data. Clearly, the mismatch of the log-energy features leads to significant performance degradation. Based on these results, we apply the proposed EMD-based method to the log-energy feature sequences in the subsequent experiments.

Table 8.2 lists the recognition accuracies of the clean-train tasks averaged over the 0-20 dB noisy test data with different degrees of speech feature post-processing. The first row in the table shows the baseline results using the raw speech features extracted by the Aurora 2.0 standard frontend. The second row in the table show the results after the application of the mean-variance normalization (MVN). MVN achieves 24.0% relative improvement. The proposed EMD-based method is applied to the log-energy feature sequences, by subtracting the first IMF for each utterance. For the MVN-feature, the relative improvement improves from 24.0% to 41.6%. The results show that the EMD-based post-processing of subtracting $c_1(t)$ from the speech feature sequence significantly reduce the mismatch between the clean and noisy feature sequences.

In addition to the basic front-end, we also apply the proposed EMD-based method on the Aurora advanced front-end (AFE). The last two rows in Table 8.2 show the result. While AFE achieves a relative improvement of 67.2% over the baseline, the application of the proposed method further improves the performance, achieving a relative improvement of 68.2% over the baseline. It is important for us to point out that AFE is a strongly robust front-end. It combines modules for voice activity detection (VAD), Wiener-filter noise reduction, and blind equalization. The fact that EMD, when applied on AFE, further improves the recognition performance is truly encouraging.

It is interesting to compare the results of subtracting different numbers of IMFs. Essentially, the more IMFs subtracted, the smoother the resultant sequence becomes. Table 8.3 lists the recognition accuracies when subtracting 1 IMFs, represented by MVN+EMD1, and 2 IMFs, represented by MVN+EMD2. From the results we can see that for the noisier data of 0 and -5 dB, MVN+EMD2 yields better accuracy. The results confirm that we should subtract fewer IMFs in higher SNRs, because the interference of noise is not as severe as in the lower-SNR cases.

From our arguments in Section 4, it is clear that the noise level and the number N of IMFs to be subtracted from the signal to reduce mismatch are closely related. Therefore, we use a scheme that allows the number of IMFs to be subtracted from the speech feature sequence to vary from utterance to utterance. We calculate the average oscillation frequency of the log-energy feature sequences from the clean-train data and use it as a threshold. If the oscillation frequency of the remainder is lower than the threshold, we stop finding and subtracting the next IMF. The results of recognition experiments are listed in Table 8.4. We can see that this scheme, denoted by MVN+EMDd, does outperform the schemes of subtracting a fixed number (1 or 2) of IMFs.

For completeness, we also study the dynamic number of IMFs to be subtracted from the AFE speech feature sequences for different SNRs. Figure 7.5 shows the sample average of N on the test set as a function of SNR for the MVN feature and the AFE feature. As expected, it increases as SNR decreases, i.e., as the noise level increases.

6 Conclusion

In this paper, we propose a feature post-processing scheme for noise-robust speech recognition frontend based on the empirical mode decomposition. We introduce EMD as a generalization of the Fourier analysis. Our motivation is that the speech generating process is non-stationary and non-linear, so EMD is theoretically superior to Fourier analysis for signal decomposition. We implement an algorithm to find the intrinsic mode functions of speech feature functions. Based on the properties of the extracted IMFs, we propose to subtract the low-order IMFs to reduce the mismatch between clean and noisy data. The evaluation results on the Aurora 2.0 database show that the proposed method can effectively improve the recognition accuracy. Furthermore, with the Aurora advanced front-end speech features, which are very noise-robust by design, the application of our proposed method further improves the recognition accuracy, which is quite remarkable.

References

1. Boll S: **Suppression of acoustic noise in speech using spectral subtraction**. *IEEE Transactions on Acoustics, Speech and Signal Processing* 1979, **27**(2):113–120.
2. Berstein A, Shallom I: **A hypothesized Wiener filtering approach to noisy speech recognition**. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* 1991:913–916.
3. Zhu W, O'Shaughnessy D: **Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm**. In *Proc. IEEE Intl. Conf. on Signal Processing* 2004:617–620.
4. Strope B, Alwan A: **A model of dynamic auditory perception and its application to robust word recognition**. *IEEE transactions on Speech and Audio Processing* 1997, **5**(5):451–464.
5. Furui S: **Cepstral analysis technique for automatic speaker verification**. *IEEE Transactions on Acoustics, Speech and Signal Processing* 1981, **29**(2):254–272.
6. Viikki O, Bye D, Laurila K: **A recursive feature vector normalization approach for robust speech recognition in noise**. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* 1998:733–736.
7. de La Torre A, Peinado A, Segura J, Perez-Cordoba J, Benitez M, Rubio A: **Histogram equalization of speech representation for robust speech recognition**. *IEEE Transactions on Speech and Audio Processing* 2005, **13**(3):355–366.
8. Huang N, Shen Z, Long S, Wu M, Shih H, Zheng Q, Yen N, Tung C, Liu H: **The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis**. *Proceeding of the Royal Society of London Series A–Mathematical Physical and Engineering Sciences* 1998, **454**:903–995.
9. Pearce D, Hirsch H: **The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions**. In *ICSA ITRW ASR2000* September 2000.
10. **Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms**. *ETSI Standard ETSI ES 202 050* 2007.

7 Figures

7.1 Figure 1

Pseudocode of the intrinsic mode function extraction.

7.2 Figure 2

Block diagram for the empirical mode decomposition analysis.

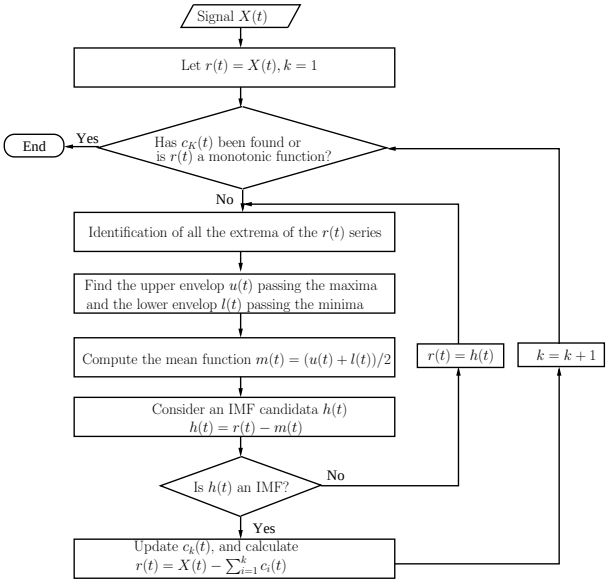
7.3 Figure 3

The intrinsic mode functions extracted from the log-energy sequence of 0dB MKG_677884ZA speech of the Aurora 2.0 database.

7.4 Figure 3

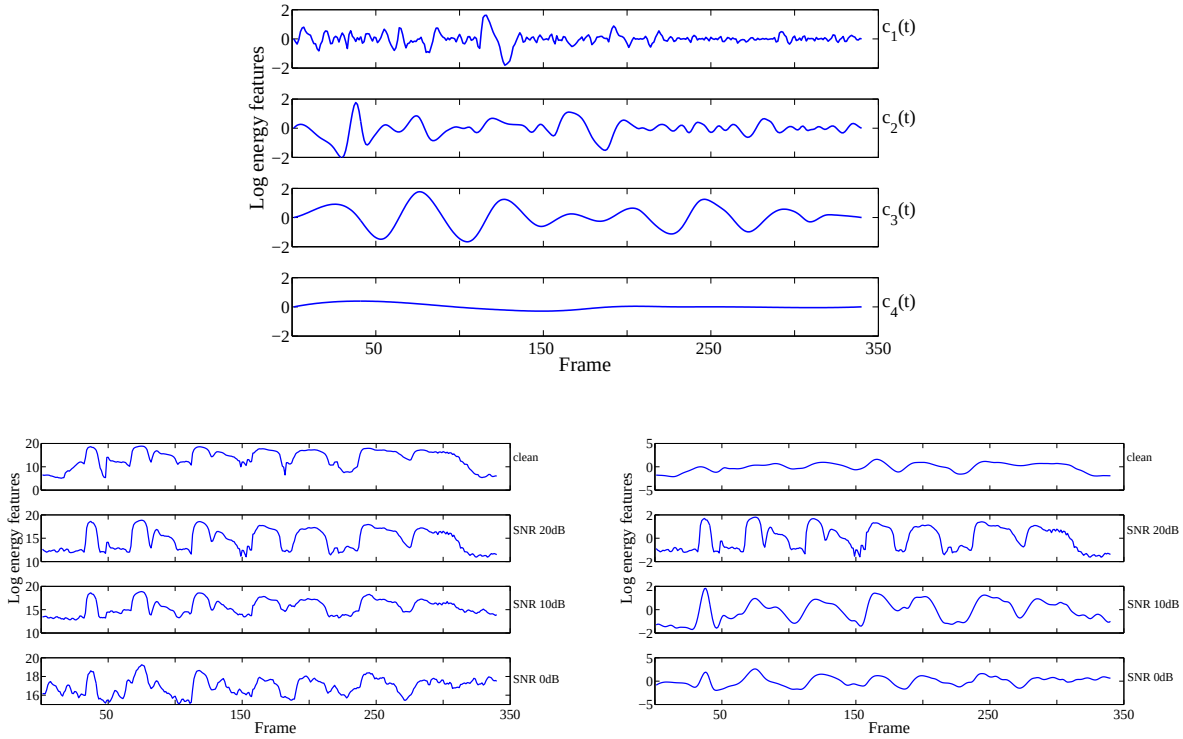
The log-energy sequences of the Aurora 2.0 utterance MKG_677884ZA under the corruption of the subway noise of different SNRs. Top: the raw log-energy sequences; Bottom: after the mean-variance normalization and the proposed EMD post-processing. Note that the different ordinate ranges in different plots.

Require: $X(t)$, K;
  $k := 1$;
  $r(t) := X(t)$;
  while $k \leq K$ $r(t)$ is not monotonic **do**
    $h(t) = 0$;
    while $h(t)$ is not an IMF **do**
      $u(t) \leftarrow$ the upper envelope of $r(t)$;
      $l(t) \leftarrow$ the lower envelope of $r(t)$;
      $h(t) \leftarrow r(t) - \frac{1}{2}\left(u(t) + l(t)\right)$;
      if ($h(t)$ is an IMF the stopping criteria is met) **then**
        $c_k(t) \leftarrow h(t)$;
        $r(t) \leftarrow X(t) - \sum_{i=1}^{k} c_i(t)$;
        $k \leftarrow k + 1$;
      else
        $r(t) \leftarrow h(t)$;
      end if
    end while
  end while
  return the IMF $c_k(t)$'s;



7.5 Figure 4

The average of the number of IMFs extracted for the MVN and AFE features as a function of SNR. For each utterance, the extraction of IMFs stops when the oscillation of $r(t)$ is below a threshold determined by the train data set.



8 Tables

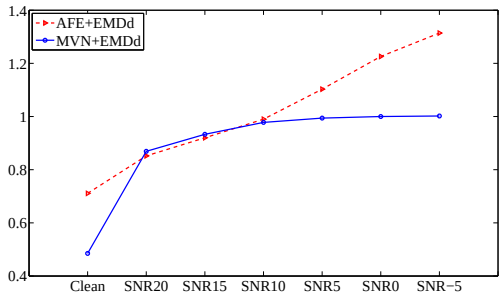
8.1 Table 1 - Performance1

Word accuracy rates of Aurora 2.0 clean-train tasks for the 0-20 dB SNR test data, replacing the noise-corrupted log energy features with clean features. The relative improvements (rel. imp.) of the word error rate over the baseline are listed.

	baseline	**replaced**	**rel. imp.**
20 dB	94.1%	98.1%	67.8%
15 dB	85.8%	96.0%	71.8%
10 dB	65.5%	90.5%	**72.5%**
5 dB	38.6%	77.6%	63.5%
0 dB	17.1%	54.7%	45.5%
0-20 dB	60.2%	83.4%	64.2%

8.2 Table 2 - Performance2

Word accuracy rates of the Aurora 2.0 clean-train tasks for the 0-20 dB SNR test data, using the proposed method.



	Set A	Set B	Set C	Avg	rel. imp.
baseline	61.3	55.8	66.1	60.1	=
MVN	70.2	70.8	66.4	69.7	24.0
MVN+EMD	76.3	77.2	76.7	**76.7**	**41.6**
AFE	87.5	87.0	85.6	86.9	67.2
AFE+EMD	88.0	87.2	86.2	**87.3**	**68.2**

8.3 Table 3 - Performance3

Word accuracy rates of Aurora 2.0 clean-train tasks for the 0-20 dB SNR test data, using 1 or 2 IMFs for subtraction.

	MVN+EMD1	MVN+EMD2	diff
clean	98.4%	98.2%	+0.2
20 dB	96.3%	96.0%	+0.3
15 dB	93.3%	92.9%	+0.4
10 dB	85.8%	85.0%	+0.8
5 dB	68.6%	68.6%	+0.0
0 dB	39.6%	41.9%	-2.3
-5 dB	16.6%	18.8%	-2.2

8.4 Table 4 - Performance4

Word accuracy rates of Aurora 2.0 clean-train tasks for the 0-20 dB SNR test data, using a dynamic number of IMFs for subtraction.

	Set A	Set B	Set C	Avg.
MVN+EMD1	76.3%	77.2%	76.6%	76.7%
MVN+EMD2	76.2%	77.5%	77.0%	76.8%
MVN+EMDd	77.6%	78.7%	77.6%	**78.0%**