A Novel Framework for Noise Robust ASR using Cochlear Implant-like Spectrally Reduced Speech ☆

Cong-Thanh Do[a,*], Dominique Pastor[b], André Goalic[b]

[a] *Idiap Research Institute, Centre du Parc, Rue Marconi 19, P.O. Box 592, CH-1920 Martigny, Switzerland*
[b] *Telecom Bretagne; UMR CNRS 3192 Lab-STICC, 29238 Brest Cedex 3, France*

Abstract

We propose a novel framework for noise robust automatic speech recognition (ASR) based on cochlear implant-like spectrally reduced speech (SRS). Two experimental protocols (EPs) are proposed in order to clarify the advantage of using SRS for noise robust ASR. These two EPs assess the SRS in both the training and testing environments. Speech enhancement was used in one of two EPs to improve the quality of testing speech. In training, SRS is synthesized from original clean speech whereas in testing, SRS is synthesized directly from noisy speech or from enhanced speech signals. The synthesized SRS is recognized with the ASR systems trained on SRS signals, with the same synthesis parameters. Experiments show that the ASR results, in terms of word accuracy, calculated with ASR systems using SRS, are significantly improved compared to the baseline non-SRS ASR systems. We propose also a measure of the training and testing mismatch based on the Kullback-Leibler divergence. The numerical results show that using the SRS in ASR systems helps in reducing significantly the training and testing mismatch due to environmental noise. The training of the HMM-based ASR systems and the recognition tests were performed by using the HTK toolkit and the Aurora 2 speech database.

Key words: Aurora 2, Cochlear implant, Kullback-Leibler divergence, HMM-based ASR, Noise robust ASR, Spectrally reduced speech.

☆This work was initially performed when Cong-Thanh Do was with Telecom Bretagne, UMR CNRS 3192 Lab-STICC. It was supported by the Bretagne regional council, Bretagne, France.

*Corresponding author: Tel +41 27 721 7764. E-mail address: cong-thanh.do@idiap.ch (C.-T. Do)

1. Introduction

Performance of automatic speech recognition (ASR) degrades significantly when environmental noise occurs during the recognition process. Actually, environmental noise is unavoidable and produces serious acoustic mismatches between the training and testing environments. Therefore, the design of ASR algorithms robust to noise, in order to reduce the mismatches between the training and testing environments, is an active research field to improve ASR performance in real operating environments.

Basically, the mismatches between training and testing environments can be reduced by (i) using noise resistant speech features and robust distance measures, e.g. (Mansour and Juang, 1989; Furui, 1986; Shannon and Paliwal, 2006; Hermansky and Morgan, 1994), (ii) transforming noisy speech into a reference environment, and recognizing noisy speech with a system trained in the reference environment, e.g. (Boll, 1979; Cooke et al., 2001; Hu and Loizou, 2003), and (iii) transforming speech models created in the reference environment in order to accommodate the noisy environment and recognize noisy speech (Leggetter and Woodland, 1995; Gales, 1998). The second approach, called speech enhancement (Gong, 1995), was primarily aimed at improving speech quality rather than improving ASR performance, since there is no direct relationship between the intended processing and the ASR performance improvement.

More specifically, in speech enhancement, attempts are made in order to transform the noisy speech into a reference environment as similar with the training environment as possible. Some typical speech enhancement methods are the spectral subtractive algorithms, e.g. (Boll, 1979; Gustafsson et al., 2001), Wiener filtering, e.g. (Loizou, 2007; Chen et al., 2008), statistical-model-based methods, e.g. (Ephraim and Malah, 1984, 1985; Hansen et al., 2006), and subspace algorithms, e.g. (Hu and Loizou, 2003; Jabloun and Champagne, 2003). In most speech enhancement algorithms, the noise spectrum estimate is assumed to be available and is essential for the satisfactory performance of speech enhancement algorithms. For instance, the noise spectrum is needed for estimating the Wiener filter or for eliminating the noise covariance matrix in the subspace algorithms (Loizou, 2007). This

noise spectrum estimate has therefore a great importance on the overall quality of the speech enhancement system. If noise is underestimated, unnatural residual noise will be perceived, whereas if noise is overestimated, enhanced speech signal will be distorted and, as a result, the speech intelligibility might be seriously affected.

Consequently, the use of a speech enhancement approach for improving ASR performance in noisy environments is more or less complex. Two main reasons can be referred. First, as mentioned above, there is no direct correlation between the efficiency of speech enhancement algorithms and the improvement of ASR performance. The enhanced speech signal might be relevant for human listener but not for the ASR system (Gong, 1995). Second, most speech enhancement algorithms need that noise spectrum be well estimated, whereas the estimation of noise spectrum is difficult, especially in presence of fluctuating or impulsive noise (Loizou, 2007).

In (Do et al., 2010a), cochlear implant-like spectrally reduced speech (SRS), which is the acoustic simulation of cochlear implant (Loizou, 1999), was shown to be relevant for hidden Markov model (HMM) based ASR using Mel frequency cepstral coefficients (MFCCs) as speech acoustic features. More specifically, the cochlear implant-like SRS, henceforth abbreviated SRS, synthesized from 16, 24, or 32 subband temporal envelopes of clean speech signal can be recognized with word accuracy (WA) (Young et al., 2006) as good as that achieved with original clean speech (Do et al., 2010a). The speech signal subband temporal envelopes, which are speech primarily temporal cues (Shannon et al., 1995), can be considered as basic information extracted from speech signal. These facts suggest that even basic information from speech signal could contain sufficient spectral information for HMM-based ASR with conventional speech feature (MFCCs). Furthermore, the use of basic (and non redundant) information in ASR might help in reducing speech variability from speech signal, especially when the speech is contaminated by environmental noise.

The purpose of this paper is to show, experimentally, that the cochlear implant-like SRS is a relevant speech model for using in ASR, and in particular, for developing ASR noise robust algorithms. More specifically, we propose to use the SRS in ASR systems where speech enhancement is used to deal with environmental noise. The SRS is assessed in both

the training and testing conditions. We propose two experimental protocols (EPs) that validate the efficiency of using SRS in ASR system in order to deal with noisy environments. In these two EPs, we compare the ASR performance of ASR systems using SRS with the standard ASR systems that do not use SRS in noisy environments. The first protocol does not involve any speech enhancement component whereas the second one involves a standard speech enhancement using a minimum mean-square error (MMSE) log-spectral amplitude estimator (Ephraim and Malah, 1985). In the second protocol, the noise spectrum is estimated by using the minima-controlled recursive averaging-2 (MCRA-2) algorithm, recently proposed by Rangachari and Loizou (Rangachari and Loizou, 2006; Loizou, 2007). In the MCRA-2 algorithm, the update of the noise estimate is faster for very rapidly varying non-stationary noise environments (Rangachari and Loizou, 2006), compared to other noise spectrum estimation algorithms, e.g. the Cohen and Berdugo's MCRA algorithm (Cohen and Berdugo, 2002). It turns out that, in noisy environment, the ASR system performance, in terms of WA, is improved when the SRS is employed, compared to the standard ASR system, which does not use SRS. On the other hand, we propose to measure the training and testing mismatch by calculating the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951) between the probability density functions (PDFs) of the speech feature vectors extracted from the training and testing speech. We show that the use of SRS in both the training and testing environments helps in reducing significantly the training and testing mismatch, due to environmental noise, measured via the KLD. The training of HMM-based ASR systems was performed by using the Aurora 2 speech database (Leonard, 1984) and the HTK speech recognition toolkit (Young et al., 2006).

2. SRS Synthesis Algorithm

A speech signal $s(t)$ is first decomposed into N subband signals $s_i(t), i = 1, \ldots, N$ by using a perceptually-motivated analysis filterbank consisting of N bandpass filters. The aim of the analysis filterbank is to simulate the motion of the basilar membrane (Kubin and Kleijn, 1999). In this respect, the filterbank consists of nonuniform bandwidth bandpass filters that are linearly spaced on the Bark scale. In this paper, each bandpass filter in the

filterbank is a second-order elliptic bandpass filter having a minimum stopband attenuation of 50dB and a 2-dB peak-to-peak ripple in the passband. The lower, upper, and central frequencies of the bandpass filters are calculated as in (Gunawan and Ambikairajah, 2004). An example of analysis filterbank is given in Fig. 1

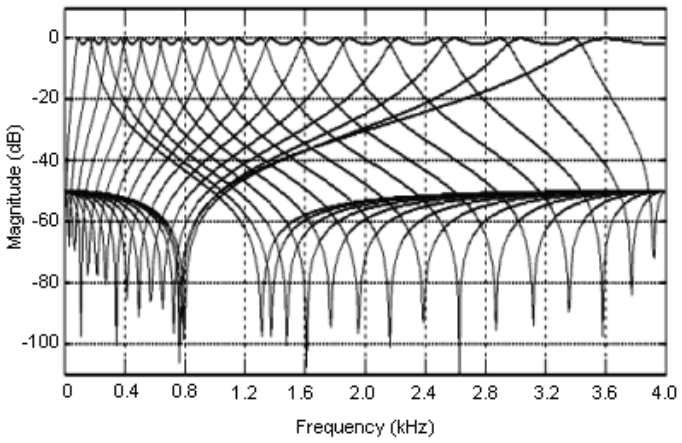

Figure 1: Frequency response of an analysis filterbank consisting of 16 second-order elliptic bandpass filters used for speech signal decomposition. The speech signal is sampled at 8 kHz.

The amplitude modulations (AMs) $m_i(t)$ of the subband signals $s_i(t), i = 1, \ldots, N$ are then extracted by, first, full-wave rectification of the outputs of the bandpass filters and, subsequently, lowpass filtering of the resulting signals. The sampling rate of the AM is kept at the same value as that of the subband signal (8 kHz). In this work, the AM filter is a fourth-order elliptic lowpass filter with 2-dB of peak-to-peak ripple and a minimum stopband attenuation of 50-dB. The subband AM $m_i(t)$ is then used to modulate a sinusoid whose frequency f_{ci} equals the central frequency of the corresponding analysis bandpass filter of that subband. Afterwards, the subband modulated signal is spectrally limited (i.e. is filtered again) by the same bandpass filter used for the original analysis subband (Shannon et al., 1995). Finally, all the subband spectrally limited signals are summed to synthesize the SRS. The mathematical formula of the SRS $\widehat{s}(t)$ can be expressed as follows, and the SRS synthesis algorithm is summarized as in Fig. 2.

$$\widehat{s}(t) = \sum_{i=1}^{N} m_i(t) \cos\left(2\pi f_{ci} t\right) \tag{1}$$

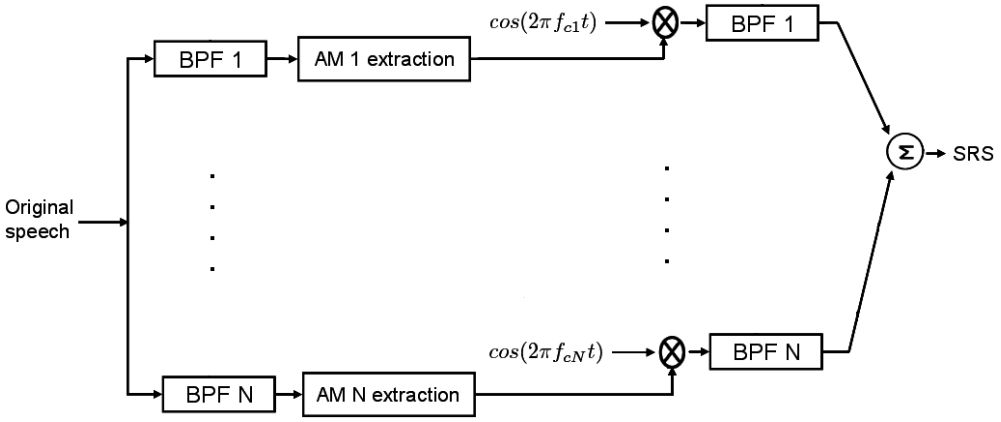

Figure 2: SRS synthesis algorithm (Shannon et al., 1995). The original speech signal is decomposed into several subband speech signals. The subband temporal envelopes are extracted from the subband speech signals and are then used to synthesize the SRS. BPF means bandpass filter.

3. Experimental Protocols (EPs)

We propose two EPs in order to validate the interest of using SRS as a speech model for noise robust ASR system. These two EPs use two different HMM-based ASR systems for the recognition of testing speech. The algorithm for SRS signal synthesis from original speech signal can be found in (Do et al., 2010a). The training of the two HMM-based ASR systems and the EPs are detailed in the following sections.

3.1. HMM-based ASR Systems Training

Two speaker-independent HMM-based ASR systems were trained on all the training part of the TI-digits speech database (Leonard, 1984) by using the HTK speech recognition toolkit (Young et al., 2006). The speech signals used for training these two systems were different. The first system was trained on the original clean speech signals of the training database of TI-digits. The second one was trained on the SRS signals synthesized from the original clean speech signals of the TI-digits training database (see Fig. 3). The two systems used bigram language models (Young et al., 2006) and their acoustic models were context-dependent three-state left-to-right triphone HMMs. The output observation distributions of the HMMs were modeled by Gaussian mixture models (GMMs) consisting of 16 Gaussian

components each. The covariance matrices of the GMMs were diagonal. The speech feature vectors consisted of 13 MFCCs that were extracted from every 25 ms length Hamming windowed speech frame by using the HTK speech recognition toolkit. The overlap between two adjacent frames was 15 ms. Furthermore, the delta and acceleration coefficients were appended to the static MFCCs to produce 39-dimensional feature vectors. In the present paper, we do not use any front-end-based noise robust ASR technique, e.g. cepstral mean subtraction, cepstral variance normalization, etc., in order to evaluate independently and thoroughly the contribution of the SRS to noise robust ASR. We designate by Mdls-Orgn the set of trained models (acoustic model, language model, and pronunciation dictionary) of the HMM-based ASR system trained on the original clean speech training database of TI-digits. In fact, "Mdls" stands for "Models" whereas "Orgn" stands for "Original". Similarly, the set of trained models of the HMM-based ASR system when this one was trained on the SRS synthesized from the clean speech signals of the TI-digits training database, is designated as Mdls-SRS.

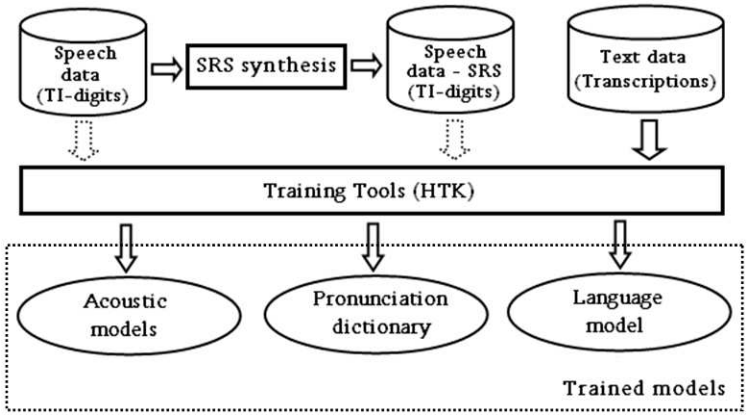

Figure 3: Training of two HMM-based ASR systems on the original clean speech signals and on the SRS, synthesized from the original clean speech signals, of the TI-digits training database, respectively. The training was performed by using the HTK speech recognition toolkit. If the speech data for training are original, the obtained models are designated by Mdls-Orgn. On the other hand, if the speech data for training are SRS synthesized from original clean speech, the obtained models are designated by Mdls-SRS.

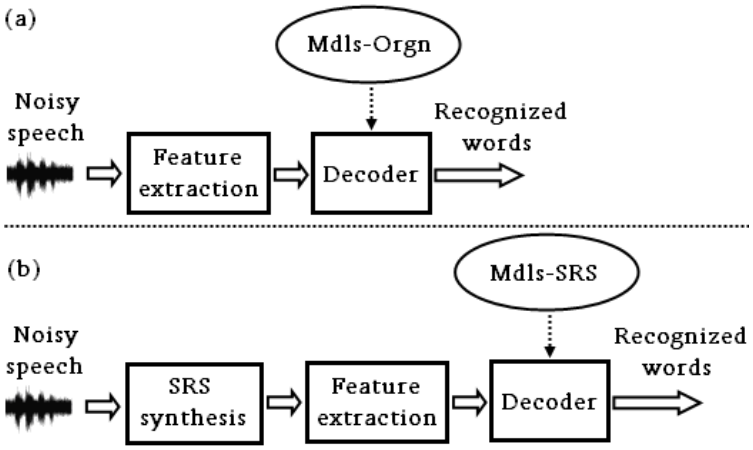

Figure 4: First experimental protocol (EP): Assessing the advantage of employing the SRS in noisy speech recognition when no speech enhancement component is used. In (a), noisy speech is directly recognized by using the HMM-based ASR system (Mdls-Orgn) trained on the original clean speech signals of the TI-digits training database. In (b), the HMM-based ASR system (Mdls-SRS) trained on the SRS signals, synthesized from the original clean speech signals in the TI-digits training database, is used to recognize the SRS signals synthesized from the noisy speech signals.

3.2. EP for noisy speech recognition

The first EP is proposed in order to assess the advantage of employing SRS in noisy speech recognition when no speech enhancement component is used. This EP is illustrated in Fig. 4. We verify whether the use of SRS signal in the ordinary condition, i.e. no speech enhancement, is beneficial or not. To this end, the Mdls-SRS is used for the recognition of the SRS that was synthesized from noisy speech. The recognition results are compared to that obtained when the Mdls-Orgn are used to recognize the original noisy speech.

3.3. EP for the recognition of enhanced noisy speech

The second EP is illustrated in Fig. 5. The purpose of this protocol is to assess the advantage of employing SRS in the recognition of noisy speech when the speech enhancement component is used to enhance noisy speech signal. We evaluate whether the synthesis of SRS from enhanced noisy speech can reduce further the variability in the enhanced speech signal. The SRS is thus synthesized from the enhanced speech signal as shown in Fig. 5. As a result, the Mdls-SRS are used for the recognition of the SRS synthesized from enhanced

speech signal whereas the Mdls-Orgn are used for the recognition of the enhanced speech signal.

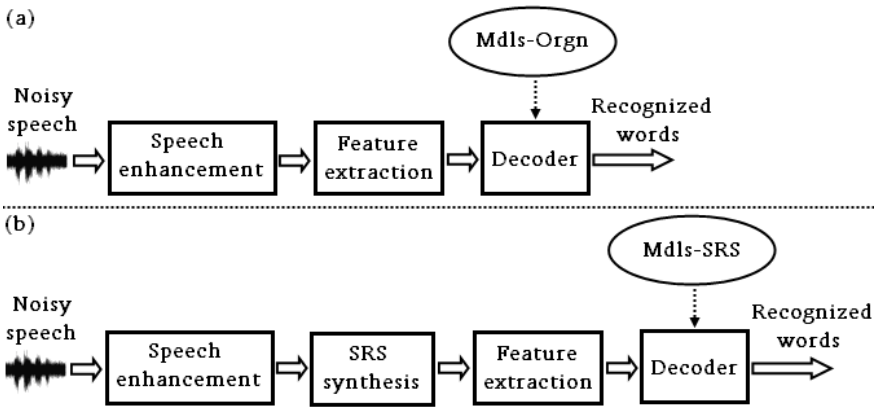

Figure 5: Second EP: Assessing the advantage of employing the SRS in the recognition of noisy speech when the speech enhancement component is used for noise reduction. As in the first EP, in (a), the HMM-based ASR system (Mdls-Orgn) trained on the original clean speech signals of the TI-digits training database is used to recognize the enhanced speech signals. On the other hand, in (b), the HMM-based ASR system (Mdls-SRS) trained on the SRS is used to recognize the SRS synthesized from the enhanced speech signals.

4. Recognition Results

4.1. Experimental Setup

4.1.1. Noisy Speech Signals for Testing

The noisy speech signals for testing are selected from the testing sets of the Aurora 2 database (Hirsch and Pearce, 2000). In this respect, 4004 utterances from 52 males and 52 females speakers in the TI-digits testing set are split into 4 subsets with 1001 utterances in each. Each subset contains the recording of all speakers. One noise signal is added to each subset of 1001 utterances at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. Furthermore, the clean case without adding noise is taken as seventh condition. Four types of noise were added into the clean speech signals in the testing sets. These four noises, including suburban train (N1), babble (N2), car (N3) and exhibition hall (N4), are used in order to investigate the performance of ASR system in the training and testing mismatch condition. In total,

these testing sets consist of 4 times 7 times 1001 = 28028 utterances, the same as in the test set A of the Aurora 2 database (Hirsch and Pearce, 2000). In addition, we use four other types of noise, restaurant (N5), street (N6), airport (N7) and train station (N8) noises, to add into the clean speech signals in the original testing sets, at 20dB, 15dB, 10dB, 5dB, 0dB and −5dB, respectively. These noisy testing sets are the same as in the test set B of the Aurora 2 database (Hirsch and Pearce, 2000). These sets consist also 4 times 7 times 1001 = 28028 utterances, in consideration of the clean case without adding noise. These testing sets are chosen in order to analyze the effectiveness of using SRS in the two proposed EPs.

4.1.2. SRS Synthesis

In the two EPs, the SRS signals for testing are synthesized from the noisy speech signals of the original testing sets. We use the SRS synthesis algorithm, as described in section 2, to synthesize the SRS signals from the original speech signals. To this end, the SRS synthesis necessitates the information concerning the analysis filterbank, the number of frequency subbands N, and the cutoff frequency of the AM filter W (or the subband temporal envelopes bandwidth). The recognition results presented in (Do et al., 2010a) showed that the synthesized SRS with $N \in \{16, 24, 32\}$ subbands and $W \in \{50, 160, 500\}$ Hz yield WAs comparable to those achieved with original clean speech. In the present paper, two types of SRS are considered from the noisy speech signals, corresponding to $(N, W) = (16, 50)$ and $(N, W) = (16, 500)$, respectively. There are thus 7 times 2 times 8 = 112 sets of SRS signals that are synthesized from the original testing sets.

4.1.3. Speech Enhancement

The speech enhancement component in the second EP is the standard MMSE log-spectral amplitude estimator (Ephraim and Malah, 1985). We have chosen such a standard speech enhancement algorithm in expecting that the obtained results could be readily generalized to other speech enhancement algorithms. The MMSE log-spectral amplitude estimator uses noise spectrum information estimated by using the MCRA-2 algorithm (Rangachari and Loizou, 2006).

Speech enhancement was applied only for testing sets of noisy speech and was not applied for clean testing speech. There are thus 6 times 8 = 48 sets of enhanced speech signals for testing (see Fig. 4). Consequently, the SRS signals are synthesized from the enhanced speech signals (see Fig. 5). The number of SRS testing sets synthesized from the testing sets of enhanced speech is 6 times 8 times 2 = 96.

4.2. Recognition Results

For the recognition experiments, and as for the training of the HMM-based ASR systems, the 39-dimensional feature vectors consisted of 13 MFCCs along with the delta and acceleration coefficients that were extracted from every 25 ms length Hamming windowed speech frame. The overlap between two adjacent frames was 15 ms. The recognition results, in terms of WA, are illustrated in Figs. 6, 7, 8 and 9.

4.2.1. First EP (speech enhancement was not used)

The WAs, calculated in the first EP with testing speech signals contaminated by four types of noise, from N1 to N4, are shown in Figs. 6(a), 6(b), 6(c) and 6(d), respectively. More specifically, Fig. 6(a) and 6(b) show the WAs calculated when the testing speech signals are contaminated by suburban train (N1) and babble (N2) noise, respectively. Similarly, Fig. 6(c) and 6(d) show the WAs calculated when the testing speech signals are contaminated by car (N3) and exhibition hall (N4) noise, respectively. In Fig. 6, WAs-Ni, $i = 1, \ldots, 4$ denote the WAs calculated in the first EP, where Mdls-Orgn are used to recognize the testing sets of noisy speech signals, contaminated by four types of noise. Similarly, the WAs-Ni-SRS(16,50) or WAs-Ni-SRS(16,500), $i = 1, \ldots, 4$, denote the WAs, calculated in the first EP, for the four types of noisy testing speech and for the two types of SRS, when the Mdls-SRS are used for recognition.

Paired t-tests revealed that WAs-Ni-SRS(16,50) are significantly greater than the WAs-Ni ($p < 0.05$), for $i = 1, \ldots, 4$. Similarly, the WAs-Ni-SRS(16,500) are significantly greater than the WAs-Ni ($p < 0.1$), for $i = 1, \ldots, 4$. Besides, the WAs-N1-SRS(16,50) is significantly smaller than the WAs-N1-SRS(16,500) ($p < 0.05$) whereas the WAs-N2-SRS(16,50) and

WAs-N3-SRS(16,50) are significantly greater than the WAs-N2-SRS(16,500) and WAs-N3-SRS(16,500), respectively ($p < 0.05$). However, no significant difference is revealed between the WAs-N4-SRS(16,50) and WAs-N4-SRS(16,500) ($p > 0.5$).

Similar interpretation could be revealed from the recognition results, measured on the testing sets contaminated by four remaining types of noise, restaurant (N5), street (N6), airport (N7) and train station (N8) noises ($i = 5, \ldots, 8$). The recognition results, in terms of WA, calculated with these testing sets are displayed in Figs. 7(a), 7(b), 7(c) and 7(d), respectively. More specifically, paired t-tests revealed that WAs-Ni-SRS(16,50) and WAs-Ni-SRS(16,500) are significantly greater than WAs-Ni ($p < 0.05$), for $i = 5, \ldots, 8$. In addition, there is no significant difference between WAs-Ni-SRS(16,50) and WAs-Ni-SRS(16,500), in terms of WA, for the four types of noise, from N5 ($i = 5$) to N8 ($i = 8$) ($p > 0.1$). In short, experimental results, performed in the first EP, with eight types of real-world noise, show that: ASR systems employing SRS outperform significantly baseline systems that do not employ SRS. In addition, in the first EP, there is no significant difference between ASR systems employing SRS(16,50) and SRS(16,500), in terms of WA improvement.

4.2.2. Second EP (standard speech enhancement was used)

In the second EP, the speech enhancement is used to enhance the quality of noisy speech signals. Then, either the Mdls-Orgn are used to recognize the enhanced speech signals or the Mdls-SRS are used to recognize the SRS synthesized from enhanced speech signals. The WAs, calculated from the recognition tests in the second EP, are shown in Fig. 8.

More specifically, Figs. 8(a), (b), (c) and (d) show the WAs, calculated from the recognition tests in which testing speech signals are contaminated by suburban train (N1), babble (N2), car (N3) and exhibition hall (N4) noise, respectively. The WAs-$\mathrm{N_E}i$, $i = 1, \ldots, 4$ denote the WAs, calculated in the second EP, where Mdls-Orgn are used to recognized the enhanced testing speech, corresponding with four types of noise. Similarly, the WAs-$\mathrm{N_E}i$-SRS(16,50) and WAs-$\mathrm{N_E}i$-SRS(16,500), $i = 1, \ldots, 4$, denote the WAs, calculated in the second EP, when the Mdls-SRS are used to recognize the corresponding SRS signals synthesized from the enhanced testing speech. In order to compare the recognition results

between speech recognition with or without speech enhancement, the WAs-Ni, $i = 1, \ldots, 4$, calculated from the first EP where speech enhancement is not used, are also displayed in Figs. 8(a), (b), (c) and (d), respectively.

Paired t-tests revealed that the WAs-$\mathrm{N_E}i$-SRS(16,500) are significantly greater than the WAs-$\mathrm{N_E}i$ for $i = 1, \ldots, 4$ ($p < 0.05$). However, no significant difference is revealed between WAs-$\mathrm{N_E}i$-SRS(16,50) and WAs-$\mathrm{N_E}i$, $i = 1, \ldots, 4$ ($p > 0.05$). For most types of noisy speech, except car noise (N3), WAs-$\mathrm{N_E}i$-SRS(16,500) are significantly greater than WAs-$\mathrm{N_E}i$-SRS(16,50) ($p < 0.05$). In addition, when the SNRs are ranged from 5 to 20dB (the high SNRs), the WAs-$\mathrm{N_E}i$-SRS(16,50) are significantly greater than the WAs-$\mathrm{N_E}i$ ($p < 0.05$), except when $i = 2$ (babble noise). On the other hand, the WAs-$\mathrm{N_E}i$, WAs-$\mathrm{N_E}i$-SRS(16,50) and WAs-$\mathrm{N_E}i$-SRS(16,500) are significantly greater than the WAs-Ni, for $i = 1, \ldots, 4$, respectively ($p < 0.05$).

As in the first EP, statistical analyses could be applied on the recognition results, measured on the four remaining types of noise, restaurant (N5), street (N6), airport (N7) and train station (N8) noises ($i = 5, \ldots, 8$). These recognition results are displayed in Figs. 9(a), (b), (c) and (d), respectively. More specifically, WAs-$\mathrm{N_E}i$, WAs-$\mathrm{N_E}i$-SRS(16,50) and WAs-$\mathrm{N_E}i$-SRS(16,500) are significantly greater than the WAs-Ni, for $i = 5, \ldots, 8$, respectively ($p < 0.05$). In addition, WAs-$\mathrm{N_E}i$-SRS(16,500) are significantly greater than WAs-$\mathrm{N_E}i$-SRS(16,50), except in case of airport noise (N7) ($p > 0.1$). Generally, in most experiments performed with the second EP, where speech enhancement was used, ASR systems employing speech enhancement + SRS outperform significantly baseline systems which employ speech enhancement only, or neither speech enhancement nor SRS. Furthermore, in most cases, ASR systems employing SRS(16,500) outperform significantly ASR systems employing SRS(16,50), in terms of WA improvement.

5. Training and Testing Mismatch Reduction

5.1. Motivation

We investigate the efficiency of using the SRS in noise robust ASR in terms of training and testing mismatch reduction. In what follows, we propose to characterize the mismatch

between the testing and training environments by the difference between the global probability density functions (PDFs) of the speech feature vectors extracted from testing and training speech signals. The speech feature vectors (MFCCs + delta + acceleration coefficients) are considered as continuously random vectors. Indeed, the difference between the PDFs of the speech feature vectors extracted from testing and training speech signals can be expected to properly characterize the mismatch between testing and training environments. The reason is that the statistics of speech feature vectors must significantly impact the recognition results in HMM-based ASR since, following Young (Young, 2008) and Nadas (Nadas, 1983), "*If speech really did have the statistics assumed by the HMMs and if there was sufficient training data, then the models estimated using maximum likelihood would be optimal in the sense of minimum variance and zeros bias*". In this section, we use the global PDFs of the speech feature vectors that were globally extracted from speech signals, not from any specific speech unit (phoneme, syllable, word, etc.), as in the HMMs.

5.2. Kullback-Leibler Divergence

There are several measures whose purpose is to characterize the difference between the probabilistic models. Such distances could be the Bhattacharyya distance (Bhattacharyya, 1943) or the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951), etc. In fact, the KLD has been widely used in the context of classification based on statistical decision theory in order to compare probabilistic models from a discrimination point of view, e.g. (Gales, 1996; Silva and Narayanan, 2006), and to globally evaluate the inherent discrimination complexity in pattern recognition (Silva and Narayanan, 2008). In this paper, we thus use the KLD to characterize the difference between the PDFs of the speech feature vectors. Let $f(\mathbf{x})$ and $g(\mathbf{x})$ be the PDFs of the speech feature vectors $\mathbf{x}$ extracted from the training and testing speech signals, respectively. The KLD $D(f(\mathbf{x})||g(\mathbf{x}))$ between $f(\mathbf{x})$ and $g(\mathbf{x})$ is defined by

$$D\left(f(\mathbf{x})||g(\mathbf{x})\right) = \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \mathrm{d}\mathbf{x} \tag{2}$$

If we assume that $f_{\mathcal{A}}$ is the global PDF of the speech feature vectors extracted from the

speech signals in the testing set $\mathcal{A}$, then $f_{\mathcal{A}}$ can be estimated, by using, for instance, the expectation-maximization (EM) algorithm (Dempster et al., 1977), on the basis of all the observations (speech feature vectors), extracted from all the speech signals in $\mathcal{A}$. Practically, the global PDFs of the speech feature vectors extracted from the training speech signals can be estimated by using the HInit tool of the HTK toolkit (Young et al., 2006). For the sake of the KLD calculation, we assume that the global PDFs of the speech feature vectors can be modeled by Gaussian mixture models (GMMs). In our case, we consider the GMMs consisting of 16 Gaussian components, with diagonal covariance matrices, as in the trained HMMs.

However, the KLD between the GMMs is not analytically tractable. Amongst the currently available methods (Hershey and Olsen, 2007) for the calculation of (2), Monte Carlo sampling is the sole method that can calculate the KLD between GMMs with arbitrary accuracy (Hershey and Olsen, 2007). We thus apply the Monte Carlo sampling method to calculate the KLDs between the speech feature vectors global PDFs that are assumed to be GMMs. More detailed about the implementation of the Monte Carlo sampling method for the calculation of the KLD can be found in (Hershey and Olsen, 2007; Do et al., 2010b).

5.3. Numerical Results

We estimate the PDFs of the speech feature vectors (MFCCs) from the speech signals in each testing set. The KLD will be calculated between the PDF of the speech feature vectors in one testing set and that of the speech feature vectors in the corresponding training set. For instance, if the testing speech is SRS (N = 16, W = 50), the KLD will be calculated between the PDF of the feature vectors extracted from testing SRS (N = 16, W = 50) and that of the feature vectors extracted from training SRS (N = 16, W = 50).

5.3.1. KLD calculation for the first EP

Fig. 10 shows the KLDs calculated between the PDFs of the speech feature vectors, extracted from the training and testing speech signals that are used in the first EP, where no speech enhancement is used (see Fig. 4). For instance, in Fig. 10(a), KLDs-N1 denotes the KLDs, calculated between the PDFs of speech feature vectors in the testing speech,

contaminated by suburban train noise (N1), at different SNRs, and that of the speech feature vectors in the clean training speech. Further, the KLDs-N1-SRS(16,50) and KLDs-N1-SRS(16,500) denote the KLDs, calculated between the PDFs of feature vectors extracted from the SRS, synthesized from noisy testing speech contaminated by suburban train noise (N1), with (N = 16, W = 50) and (N = 16, W = 500), respectively, and that of the feature vectors, extracted from the corresponding SRS for training. The curves in Figs 10(b), (c) and (d) have the similar meanings but for the three other noises, babble (N2), car (N3) and exhibition hall (N4).

Paired t-tests revealed that KLDs-Ni-SRS(16,50) are significantly smaller than KLDs-Ni, for $i = 1, \ldots, 4$ ($p < 0.01$). Similarly, KLDs-Ni-SRS(16,500) are significantly smaller than KLDs-Ni ($p < 0.05$), except when $i = 2$, babble noise ($p > 0.2$). On the other hand, KLDs-Ni-SRS(16,50) are significantly smaller than KLDs-Ni-SRS(16,500) ($p < 0.05$), except when the contaminated noise is babble (N2).

Numerical results of KLDs calculations with the testing speech signals, contaminated by four remaining types of noise, restaurant (N5), street (N6), airport (N7), and train station (N8), are displayed in Figs. 11(a), 11(b), 11(c) and 11(d), respectively. As in the previous analyses, paired t-tests revealed that KLDs-Ni-SRS(16,50) and KLDs-Ni-SRS(16,500) are significantly smaller than KLDs-Ni, for $i = 5, \ldots, 8$ ($p < 0.01$). On the other hand, KLDs-Ni-SRS(16,50) are significantly smaller than KLDs-Ni-SRS(16,500) ($p < 0.05$), for four types of noise, from N5 to N8.

5.3.2. KLD calculation for the second EP

The KLDs, calculated between the PDFs of the speech feature vectors extracted from the training and testing speech signals, which are used in the second EP, are shown in Fig. 12. In the second EP, standard speech enhancement is used and the recognition experiments are performed on the enhanced speech signals or SRS synthesized from the enhanced speech signals. In Fig. 12(a), (b), (c) and (d), KLDs-N$_{\mathrm{E}}i$, $i = 1, \ldots, 4$ denote the KLDs, calculated between the PDFs of speech feature vectors extracted from enhanced testing speech, contaminated by suburban train (N1), babble (N2), car (N3) and exhibition hall (N4) noise,

respectively, and that of the speech feature vectors extracted from clean training speech. Similarly, the KLDs-$N_E i$-SRS(16,50) and KLDs-$N_E i$-SRS(16,500), $i = 1, \ldots, 4$, denote the KLDs calculated between the PDFs of speech feature vectors in the testing speech, which is SRS synthesized from enhanced speech signals, and those of the speech feature vectors in the corresponding SRS for training. Similarly, Fig. 13(a), (b), (c) and (d) display the KLDs, calculated between the PDFs of the speech feature vectors extracted from the training and testing speech signals, contaminated by restaurant (N5), street (N6), airport (N7) and train station (N8) noises, respectively.

Paired t-tests are performed and reveal that the KLDs-$N_E i$-SRS(16,50) and KLDs-$N_E i$-SRS(16,500) are significantly smaller than the KLD-$N_E i$, $i = 1, \ldots, 4$ ($p < 0.05$). Furthermore, KLDs-$N_E i$-SRS(16,50) are significantly smaller than KLDs-$N_E i$-SRS(16,500), $i = 1, \ldots, 4$ ($p < 0.05$). For the four remaining types of noise, from N5 ($i = 5$) to N8 ($i = 8$), similar conclusions could be revealed through the statistical tests. That is, KLDs-$N_E i$-SRS(16,50) and KLDs-$N_E i$-SRS(16,500) are significantly smaller than the KLD-$N_E i$, $i = 5, \ldots, 8$ ($p < 0.001$). In addition, KLDs-$N_E i$-SRS(16,50) are significantly smaller than KLDs-$N_E i$-SRS(16,500), $i = 5, \ldots, 8$ ($p < 0.05$).

6. Conclusion and Perspective

6.1. Conclusion

In this paper, we have introduced a novel framework for noise robust ASR based on cochlear implant-like SRS (Do et al., 2010a). Two experimental protocols have been proposed in order to emphasize the relevance of SRS for noise robust ASR. These two EPs assess SRS in both the training and testing environments. The second EP employs speech enhancement component whereas the first one does not. In training, the SRS is synthesized from original clean speech whereas in testing, the SRS is synthesized directly from noisy speech or from enhanced speech signals. The synthesized SRS has then been recognized with the models trained on the corresponding SRS signals. Experiments have shown that the ASR results, in terms of WA, calculated by using the two proposed EPs, are signifi-

cantly improved in comparison with the baseline non-SRS ASR systems trained on TI-digits database.

More specifically, according to the first EP results, when no speech enhancement is used, using SRS ($N = 16, W = 50$) and ($N = 16, W = 500$) in both the training and testing environments help in improving significantly the WA when the testing speech signals are contaminated by real-world noises, namely suburban train, babble, car, exhibition hall, restaurant, street, airport and train station noises. From the results of the second EP, it follows that when a standard MMSE log-spectral amplitude estimator (Ephraim and Malah, 1985) speech enhancement is used, WA is significantly improved when the Mdls-SRS ($N = 16, W = 500$) are used to recognize the SRS ($N = 16, W = 500$) synthesized from enhanced speech signals, in case of the testing speech signals are contaminated by suburban train, babble, car, exhibition hall, restaurant, street, airport and train station noises. On the other hand, the WAs obtained in the second EP, where speech enhancement is used, are significantly better than those obtained in the first EP, where no speech enhancement and no SRS are used. It also follows from the results of the second EP that, when the testing speech signals are contaminated by noises (suburban train, car and exhibition hall), at the SNRs ranging from 5dB to 20dB, the ASR systems using SRS ($N = 16, W = 50$) in both the training and testing environments can improve significantly the WAs. For the four types of noise, restaurant, street, airport and train station noises, ASR systems employing SRS ($N = 16, W = 50$) also make significant WA improvement for all SNRs, from -5dB to 20dB. These results show that the SRS, with a broader bandwidth of the subband temporal envelopes, can make it possible to improve the WAs in the two EPs, either with or without speech enhancement components, in a broader range of SNRs and type of noises. These results highlights also the strength of speech enhancement in ASR as well as the role of SRS as an efficient complement for noise robust ASR systems based on speech enhancement. As a result, the SRS ($N = 16, W = 500$) is recommended to use in noise robust ASR systems, based on speech enhancement, in order to gain better WA improvement. However, further analyses should be carried out in order to assess faithfully this conclusion as well as the reason of the phenomenon.

We have also proposed a measure of the training and testing mismatch. This measure is based on the Kullback-Leibler divergence. The KLDs are calculated between the PDFs of the speech feature vectors extracted from the training and testing speech signals. This train/test mismatch measure is relevant since the PDFs of the speech feature vectors would impact significantly the recognition results of HMM-based ASR systems. In the first EP, when the speech signals are contaminated by suburban train, car, exhibition hall, restaurant, street, airport and train station noises, the numerical results have shown that using the SRS ($N = 16, W = 50$ and $N = 16, W = 500$), in both the training and testing environments, helps in reducing significantly the training and testing mismatches measured via the KLDs. These significant reductions were also observed in the second EP for all types of noise (suburban train, babble, car, exhibition hall, restaurant, street, airport and train station). In addition, the reduction of training and testing mismatches, measured via the KLDs, attained by using the SRS ($N = 16, W = 50$), is significantly better than that obtained with the SRS ($N = 16, W = 500$), for all the types of noise in the two proposed EPs, except in the first EP when the testing speech signals are contaminated by babble noise. The training and testing mismatch measure via the KLD is, however, only a "coarse" measure and does not make it possible to predict precisely the ASR performance in terms of WA. Actually, the training and testing mismatch analyses, based on the KLDs, and the recognition results, in terms of WA, are not completely unified. Nevertheless, the KLD makes it possible to anticipate the tendency of the training and testing mismatch reduction, and to emphasize the role of the SRS in reducing the training and testing mismatch due to environmental noise. In this respect, the SRS and the training and testing mismatch measure, based on the KLD, contribute to the constitution of a unified framework for noise robust ASR.

6.2. Perspective

Several prospects can be derived from the results presented in this paper. In the present work, two types of SRS, corresponding to ($N = 16, W = 50$) and ($N = 16, W = 500$), are used for the experiments. More research work can be performed in order to find the SRS parameters that yield the best recognition results with the two proposed EPs. The fact

that ASR system performance is improved when the acoustic models are triphone HMMs suggests potential application of the SRS in large-vocabulary ASR systems (Gauvain and Lamel, 2000), trained on others speech databases, in order to get better performance in noisy conditions. From the signal processing point of view, the presented results show that the cochlear implant-like SRS is a relevant speech model for using in ASR, and in particular, for developing ASR noise robust algorithms. Indeed, the calculations in these algorithms, e.g. in the MMSE log-spectral amplitude estimator (Ephraim and Malah, 1985), might be adapted to the SRS model in order to achieve better speech enhancement or noise reduction performance. Furthermore, these algorithms might operate in the broadband or in the frequency subbands of the SRS. On the other hand, the successful measure of the training and testing mismatch, based on the KLD (or even other distances), should introduce new prospects on ASR performance optimization based on this divergence. Indeed, this measure might be used as a cost function so as its decrease should entail the increase of ASR system performance. However, further reflection and experiments are needed to assess the feasibility of this idea.

Acknowledgement

The authors are grateful to the reviewers for their relevant remarks that help in improving the quality of the manuscript. They would also like to thank Philip N. Garner (Idiap Research Institute, Switzerland) for his ASR training scripts and the useful discussions around the Aurora 2 database.

References

D. Mansour and B.-H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 37, no. 11, pp. 1659-1671, Nov. 1989.

S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 32, no. 4, pp. 357-366, Aug. 1980.

B. J. Shannon and K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Communication*, Vol. 48, no. 11, pp. 1458-1485, Nov. 2006.

H. Hermansky and N. Morgan "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, no. 4, pp. 578-589, Oct. 1994.

S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 27, no. 2, pp. 113-120, Apr. 1979.

Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. on Speech and Audio Processing*, Vol. 11, no. 4, pp. 334-341, Jul. 2003.

C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, no. 2, pp. 171-185, Apr. 1995.

M. J. F. Gales, "Predictive model-based compensation schemes for robust speech recognition," *Speech Communication*, Vol. 25, no. 1-3, pp. 49-74, Aug. 1998.

Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, Vol. 16, no. 3, pp. 261-291, Apr. 1995.

H. Gustafsson, S. Nordholm and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE on Speech and Audio Processing*, Vol. 9, no. 8, pp. 799-807, Nov. 2001.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, Vol. 34, no. 3, pp. 267-285, Jun. 2001.

P. Loizou, "Speech enhancement: theory and practice," *CRC: Boca Raton, FL*, 2007.

J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," *Springer handbook of speech processing*, (J. Benesty, M. M. Sondhi, and Y. Huang Eds. Springer, pp 843-871, 2008).

Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 33, no. 2, pp. 443-445, Apr. 1985.

J. H. L. Hansen, V. Radhakrishnan, and K. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, no. 6, pp. 2049-2063, Nov. 2006.

F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, Vol. 11, no. 6, pp. 700-708, Nov. 2003.

C.-T. Do, D. Pastor and A. Goalic, "On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 18,

no. 5, pp. 1065-10688, Jul. 2010.

P. Loizou, "Introduction to cochlear implants," *IEEE Engineering in Medecine and Biology Magazine*, Vol. 18, no. 1, pp. 32-42, Jan-Feb. 1999.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollarson, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," *Cambridge university engineering department*, 2006.

R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, Vol. 270, no. 5234, pp. 303-304, Oct. 1995.

S. Rangachari and P. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, Vol. 48, no. 2, pp. 220-231, Feb. 2006.

I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, Vol. 9, no. 1, pp. 12-15, Jan. 2002.

S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, no. 1, pp. 79-86, 1951.

R. Leonard, "A database for speaker-independent digit recognition," *Proc. IEEE ICASSP 1984, March 19 - 21, San Diego, USA*, Vol. 9, pp. 328-331, 1984.

S. Quackenbush, T. Barnwell, and M. Clements, "Objective measures of speech quality," *Englewood Cliffs, NJ, Prentice-Hall*, 1988.

ITU-T, "Objective measurement of active speech level," *ITU-T Recommendation*, 1993, p. 56.

A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, Vol. 12, no. 3, pp. 247-251, Jul. 1993.

J. H. McDonald, "Handbook of Biological Statistics, 2nd Eds.," *Baltimore, Maryland, Sparky House Publishing*, 2009.

S. Young, "HMMs and related speech technologies," *Springer handbook of speech processing*, (J. Benesty, M. M. Sondhi, and Y. Huang Eds. Springer, pp 539-557, 2008).

A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 31, no. 4, pp. 814-817, Aug. 1983.

A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, Vol. 35, pp. 99-109, 1943.

M. Gales, "Model-based techniques for noise robust speech recognition," *PhD Thesis*, Cambridge University, 1996.

J. Silva and S. Narayanan, "Average divergence distance as a statistical discrimination measure for hidden

Markov models," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 14, no. 3, pp. 890-906, May. 2006.

J. Silva and S. Narayanan, "Upper bound Kullback-Leibler divergence for transient hidden Markov models," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 56, no. 9, pp. 4176-4188, Sep. 2008.

A. P. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, Vol. 39, no. 1, pp. 1-38, 1977.

J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," *Proc. IEEE ICASSP 2007, April 15 - 20, Hawaii, USA*, Vol. 4, pp. 317-324, 2007.

C.-T. Do, D. Pastor, and A. Goalic, "Corrélation entre les différences entre les taux de reconnaissance de la parole sur deux ensembles de test et celles des distributions de probabilité des vecteurs acoustiques de ces même ensembles," *Proc. JEP 2010 - Journées d'Etude sur la Parole, May 25 - 28, Mons, Belgium*, pp. 49-52, 2010.

J.-L. Gauvain and L. Lamel, "Large-vocabulary continuous speech recognition: advances and applications," *Proc. IEEE*, Vol. 88, no. 8, pp. 1181-1200, Aug. 2000.

G. Kubin and W. B. Kleijn, "On speech coding in a perceptual domain," *Proc. IEEE ICASSP 1999, March 15 - 19, Phoenix, AZ, USA*, Vol. 1, pp. 205-208, 1999.

T. S. Gunawan and E. Ambikairajah, "Speech enhancement using temporal masking and fractional Bark gammatone filters," *Proc. 10th Australian International Conference on Speech Science & Technology, December 8 - 10, Sydney, Australia*, Dec, pp. 420-425, 2004.

H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ISCA ASR2000: Automatic speech recognition: Challenges for the new millenium, September 18 - 20, Paris, France*, 2000.

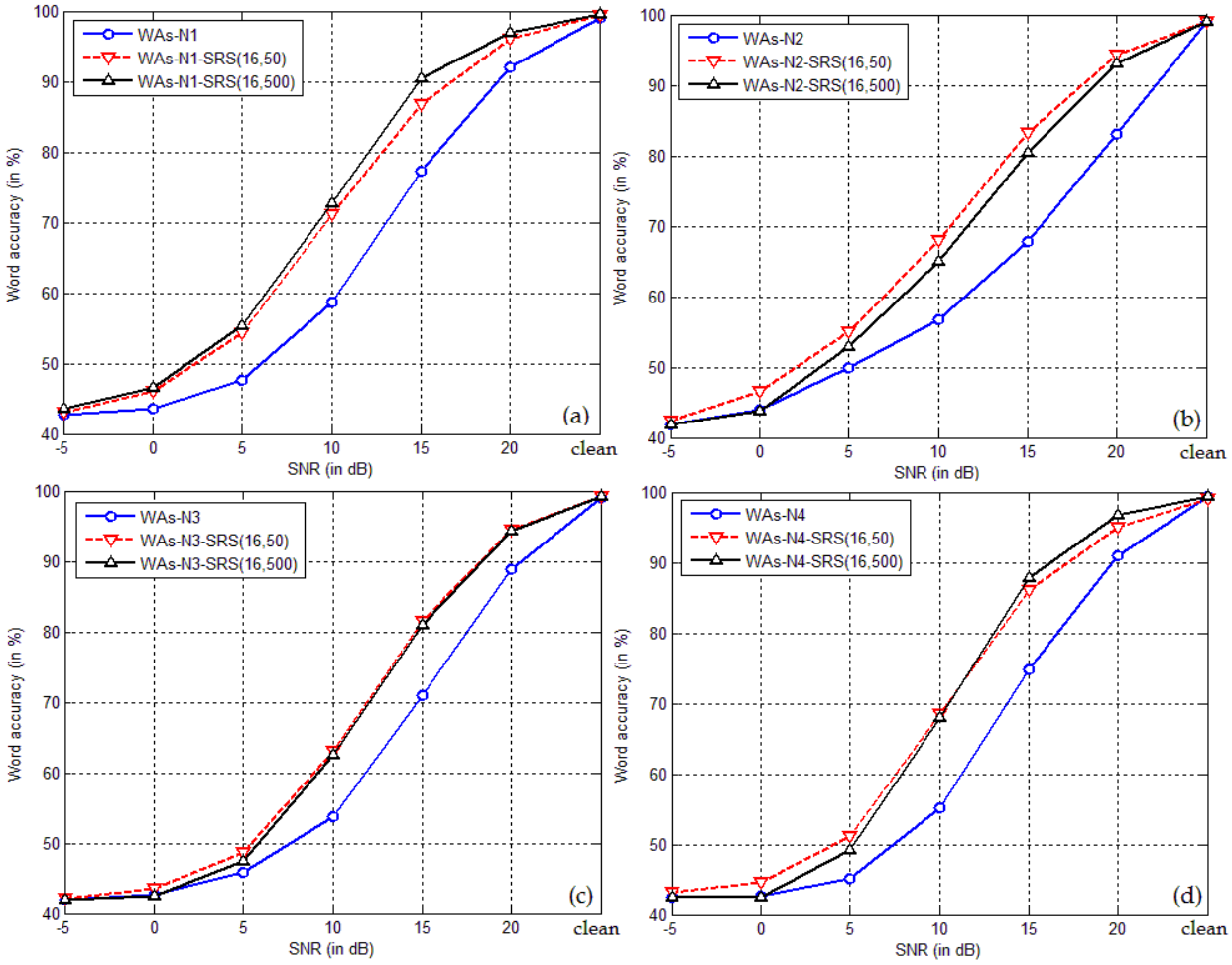

Figure 6: Recognition results, in terms of word accuracy, obtained from the first EP. Fig. 6(a) and 6(b) display the WAs calculated when the testing speech signals are contaminated by suburban train (N1) and babble (N2) noises, respectively. Similarly, Fig. 6(c) and 6(d) display the WAs calculated when the testing speech signals are contaminated by car (N3) and exhibition hall (N4) noises, respectively. For instance, in Fig. 6(a), the WAs-N1 represents the WAs obtained from the recognition of speech signals, contaminated by suburban train noise, using the Mdls-Orgn. Similarly, in Fig. 6(b), the WAs-N2-SRS$(16, 50)$ and WAs-N2-SRS$(16, 500)$ represent the WAs obtained from the recognition of the SRS synthesized from babble noise contaminated speech signals by using the Mdls-SRS, when the SRS synthesis parameters (N, W) equal $(16, 50)$ and $(16, 500)$, respectively.

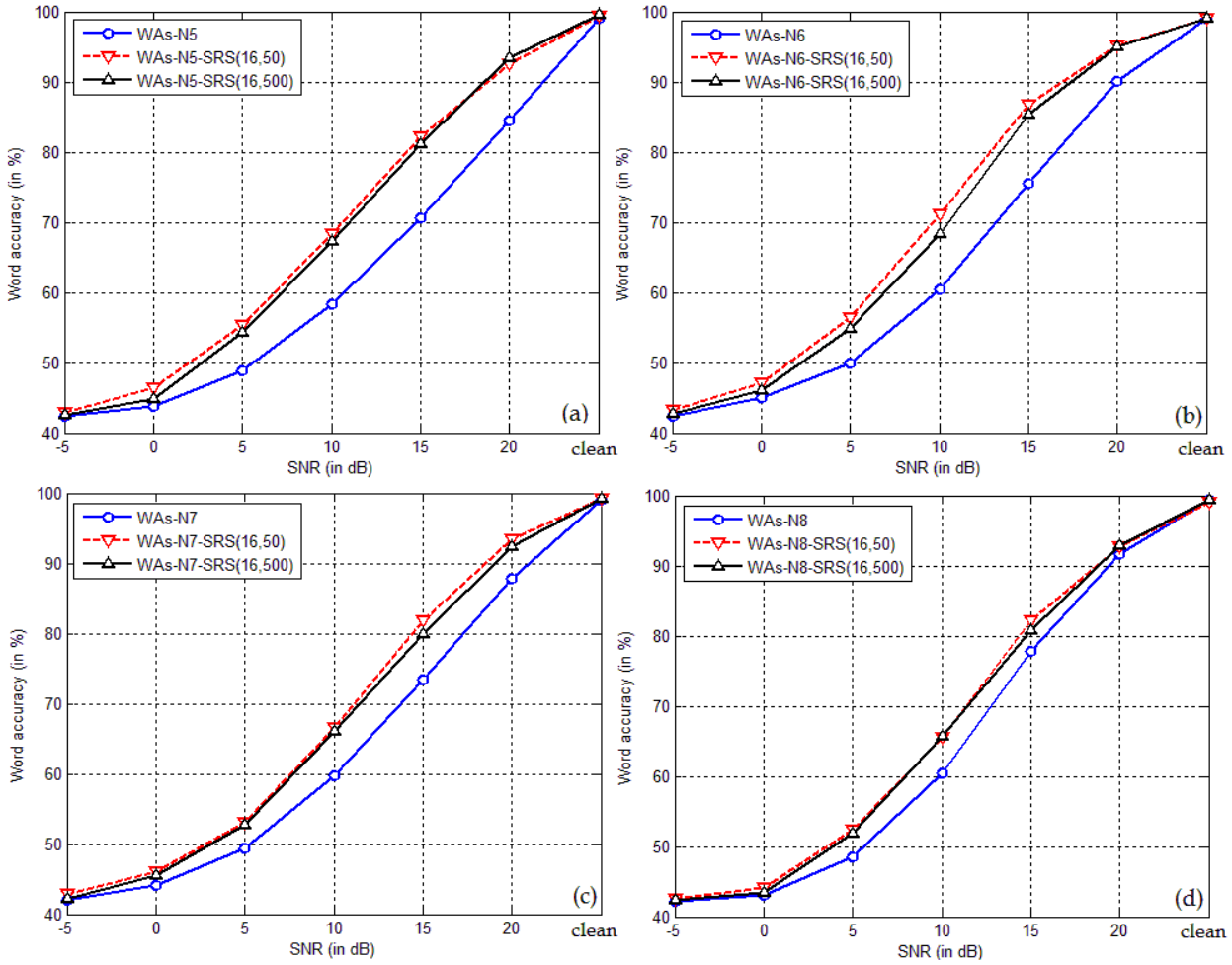

Figure 7: Recognition results, in terms of word accuracy, obtained from the first EP. Fig. 7(a) and 7(b) display the WAs calculated when the testing speech signals are contaminated by restaurant (N5) and street (N6) noises, respectively. Similarly, Fig. 7(c) and 7(d) display the WAs calculated when the testing speech signals are contaminated by airport (N7) and train station (N8) noises, respectively. For instance, in Fig. 7(a), the WAs-N5 represents the WAs obtained from the recognition of speech signals, contaminated by restaurant noise, using the Mdls-Orgn. Similarly, in Fig. 7(b), the WAs-N6-SRS$(16, 50)$ and WAs-N6-SRS$(16, 500)$ represent the WAs obtained from the recognition of the SRS synthesized from street noise contaminated speech signals by using the Mdls-SRS, when the SRS synthesis parameters (N, W) equal $(16, 50)$ and $(16, 500)$, respectively.

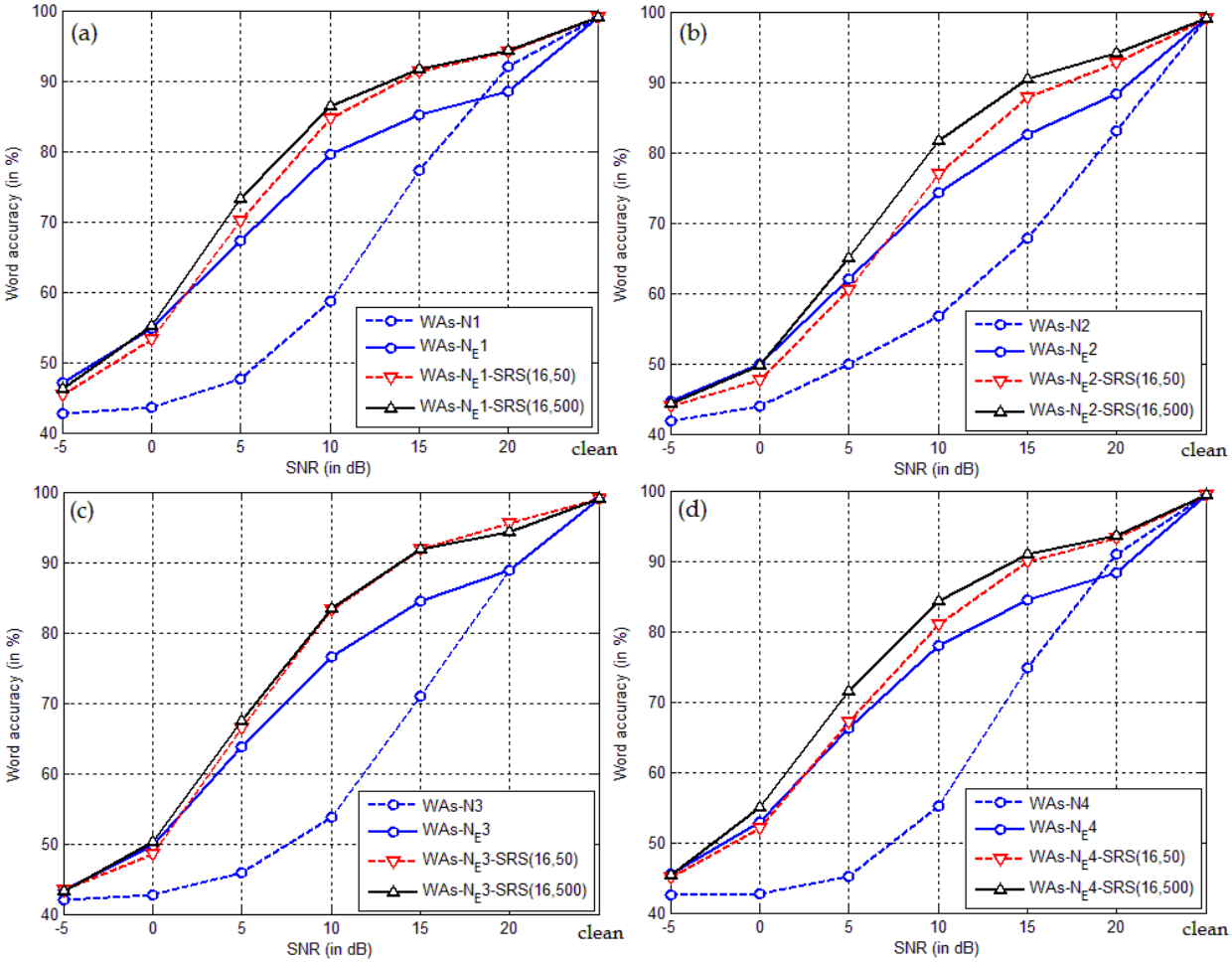

Figure 8: Recognition results, in terms of word accuracy, obtained from the second EP, where the standard speech enhancement was used. The WAs calculated in the fisrt EP, in case of no speech enhancement and no SRS are used, WAs-Ni, $i = 1, \ldots, 4$, are also displayed. Fig. 8(a) and (b) display the WAs issued from the recognition of the enhanced speech signals that were contaminated by suburban train (N1) and babble (N2) noise, respectively. Similarly, Fig. 8(c) and (d) display the WAs issued from the recognition of the enhanced speech signals that were contaminated by car (N3) and exhibition all (N4) noise, respectively. For instance, in Fig. 8(c), the WAs-$\mathrm{N_E}3$ represents the WAs obtained from the recognition of the enhanced speech signals, that were contaminated by car noise, using the Mdls-Orgn. Similarly, in Fig. 8(d), the WAs-$\mathrm{N_E}4$-SRS$(16, 50)$ and the WAs-$\mathrm{N_E}4$-SRS$(16, 500)$ represent the WAs obtained from the recognition of the SRS synthesized from the enhanced speech signals, that were contaminated by exhibition hall noise, by using the Mdls-SRS, when the SRS synthesis parameters (N, W) equal $(16, 50)$ and $(16, 500)$, respectively.

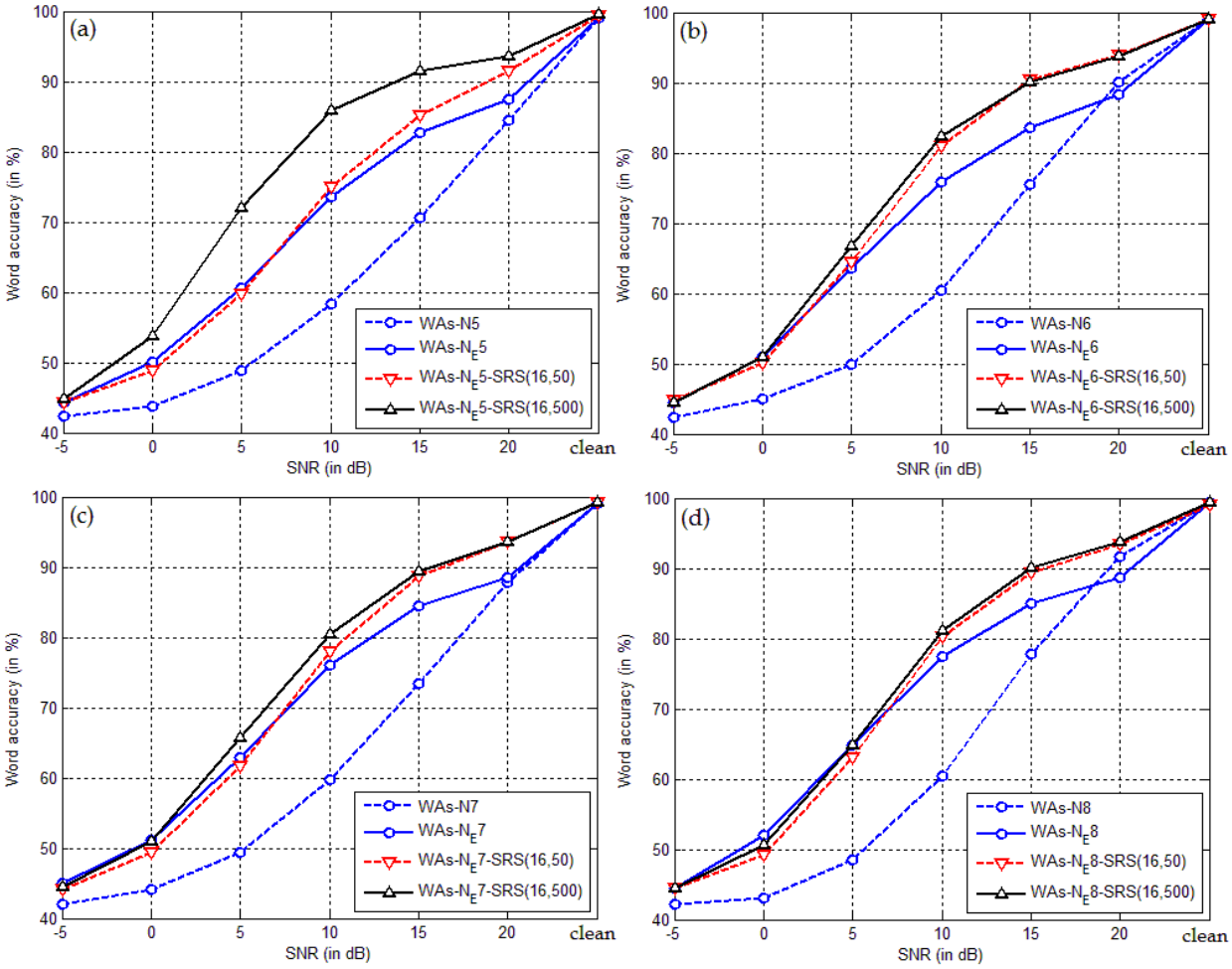

Figure 9: Recognition results, in terms of word accuracy, obtained from the second EP, where the standard speech enhancement was used. The WAs calculated in the fisrt EP, in case of no speech enhancement and no SRS are used, WAs-Ni, $i = 5, \ldots, 8$, are also displayed. Fig. 9(a) and (b) display the WAs issued from the recognition of the enhanced speech signals that were contaminated by restaurant (N5) and street (N6) noises, respectively. Similarly, Fig. 9(c) and (d) display the WAs issued from the recognition of the enhanced speech signals that were contaminated by airport (N7) and train station (N8) noises, respectively. For instance, in Fig. 9(c), the WAs-$\mathrm{N_E}7$ represents the WAs obtained from the recognition of the enhanced speech signals, that were contaminated by airport noise, using the Mdls-Orgn. Similarly, in Fig. 9(d), the WAs-$\mathrm{N_E}8$-SRS$(16, 50)$ and the WAs-$\mathrm{N_E}8$-SRS$(16, 500)$ represent the WAs obtained from the recognition of the SRS synthesized from the enhanced speech signals, that were contaminated by train station noise, by using the Mdls-SRS, when the SRS synthesis parameters (N, W) equal $(16, 50)$ and $(16, 500)$, respectively.

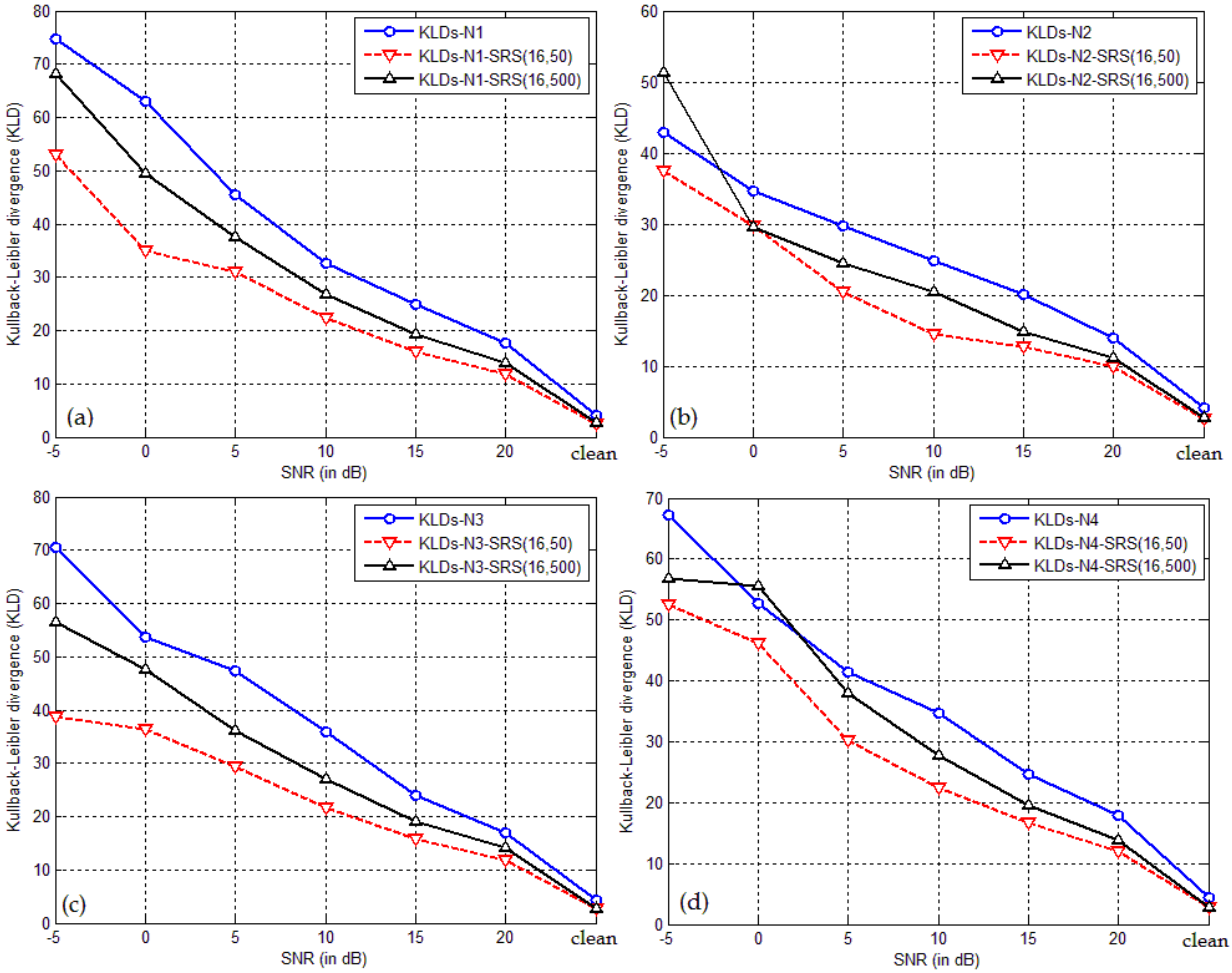

Figure 10: Kullback-Leibler divergences (KLDs) calculated between the PDFs of the speech feature vectors extracted from the testing and training speech signals. Each PDF was estimated from all the observations extracted from all the speech signals in the corresponding testing set that was used in the first EP, where no speech enhancement was used. Fig. 10(a), (b), (c) and (d) display the KLDs calculated in the context that the speech signals for testing are contaminated by suburban train (N1), babble (N2), car (N3) and exhibition hall (N4) noises, respectively. For instance, in Fig. 10(a), KLDs-N1 denotes the KLDs calculated between the PDFs of speech feature vectors in the testing speech, contaminated by suburban train noise (N1), and that of the feature vectors in the clean training speech. Further, the KLDs-N1-SRS(16,50) and KLDs-N1-SRS(16,500) denote the KLDs calculated between the PDFs of feature vectors extracted from the SRS in the testing sets, with (N = 16, W = 50) and (N = 16, W = 500), respectively, and that of the feature vectors extracted from the corresponding SRS for training.

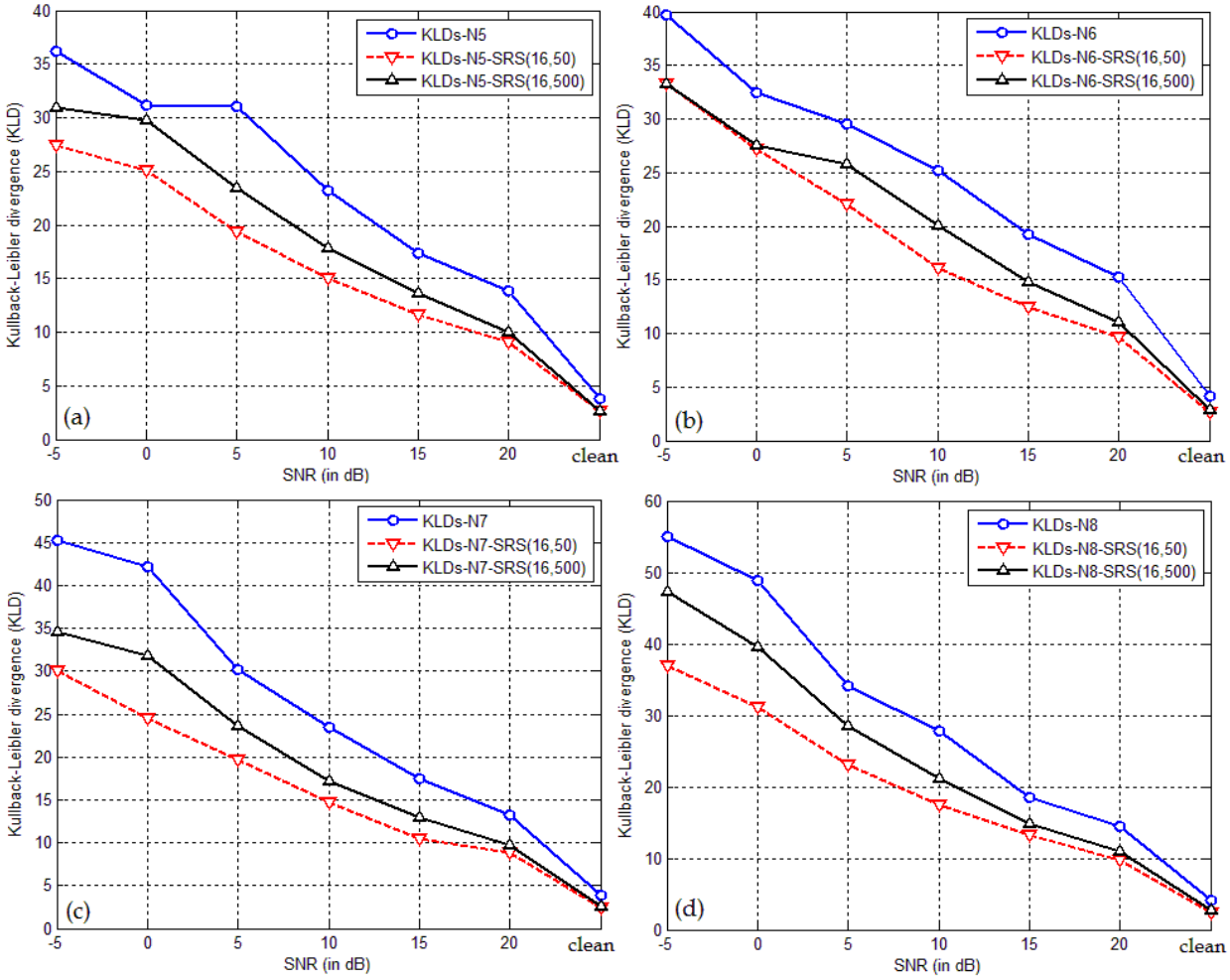

Figure 11: Kullback-Leibler divergences (KLDs) calculated between the PDFs of the speech feature vectors extracted from the testing and training speech signals. Each PDF was estimated from all the observations extracted from all the speech signals in the corresponding testing set that was used in the first EP, where no speech enhancement was used. Fig. 11(a), (b), (c) and (d) display the KLDs calculated in the context that the speech signals for testing are contaminated by restaurant (N5), street (N6), airport (N7) and train station (N8) noises, respectively. For instance, in Fig. 11(a), KLDs-N5 denotes the KLDs calculated between the PDFs of speech feature vectors in the testing speech, contaminated by restaurant noise (N5), and that of the feature vectors in the clean training speech. Further, the KLDs-N5-SRS(16,50) and KLDs-N5-SRS(16,500) denote the KLDs calculated between the PDFs of feature vectors extracted from the SRS in the testing sets, with (N = 16, W = 50) and (N = 16, W = 500), respectively, and that of the feature vectors extracted from the corresponding SRS for training.

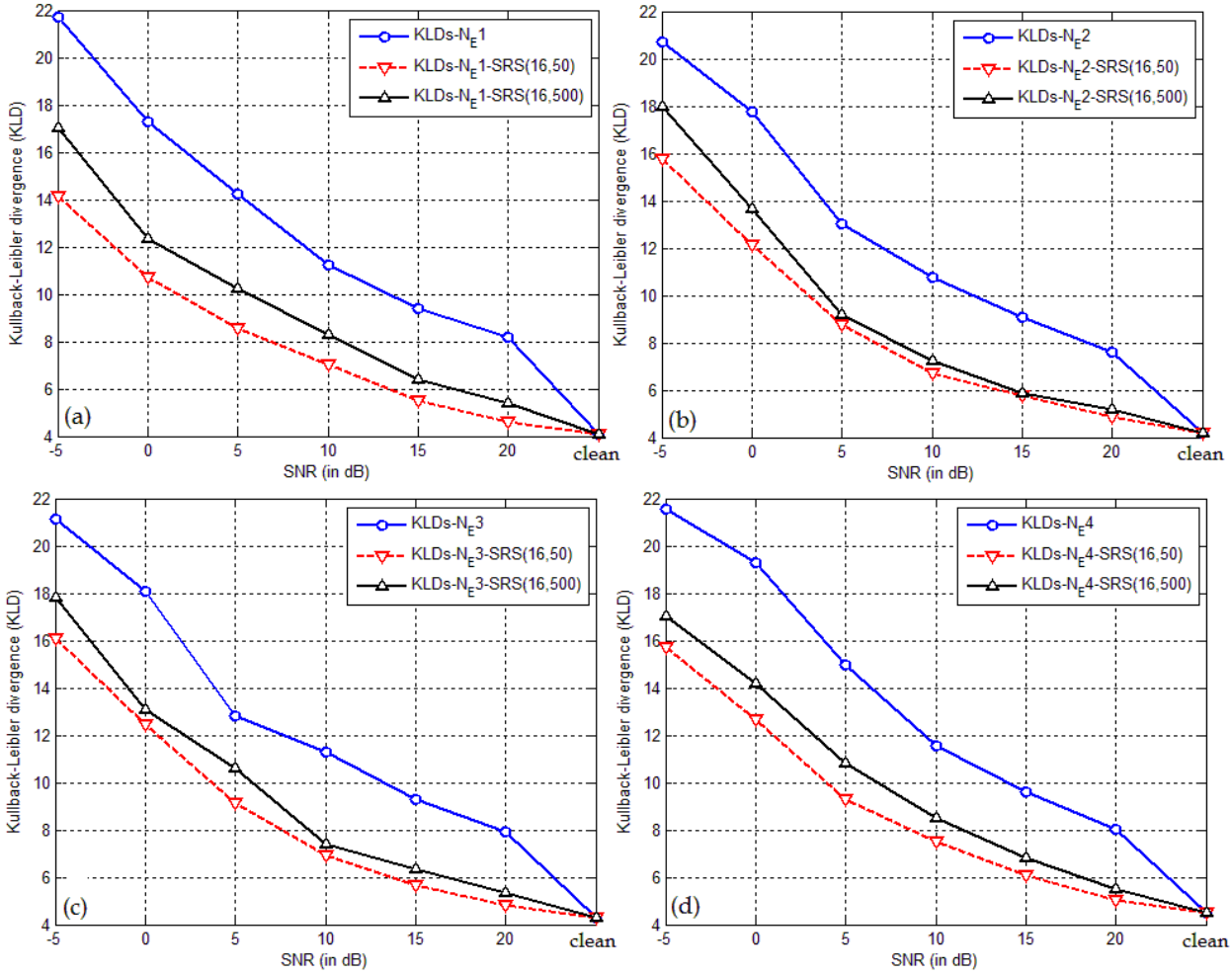

Figure 12: KLDs calculated between the PDFs of the speech feature vectors extracted from the testing and training speech signals. Each PDF was estimated from all the observations extracted from all the speech signals in the corresponding testing set that was used in the second EP, where the speech enhancement component was used. In Fig. 12(a), (b), (c) and (d), KLDs-N$_E i$, $i = 1, \ldots, 4$ denote the KLDs calculated between the PDFs of speech feature vectors extracted from testing speech, contaminated by suburban train (N1), babble (N2), car (N3) and exhibition hall (N4) noises, respectively, and that of the speech feature vectors extracted from clean training speech. Similarly, the KLDs-N$_E i$-SRS(16,50) and KLDs-N$_E i$-SRS(16,500), $i = 1, \ldots, 4$, denote the KLDs calculated between the PDFs of speech feature vectors in the testing speech, which is SRS synthesized from enhanced speech signals, and the corresponding SRS for training.

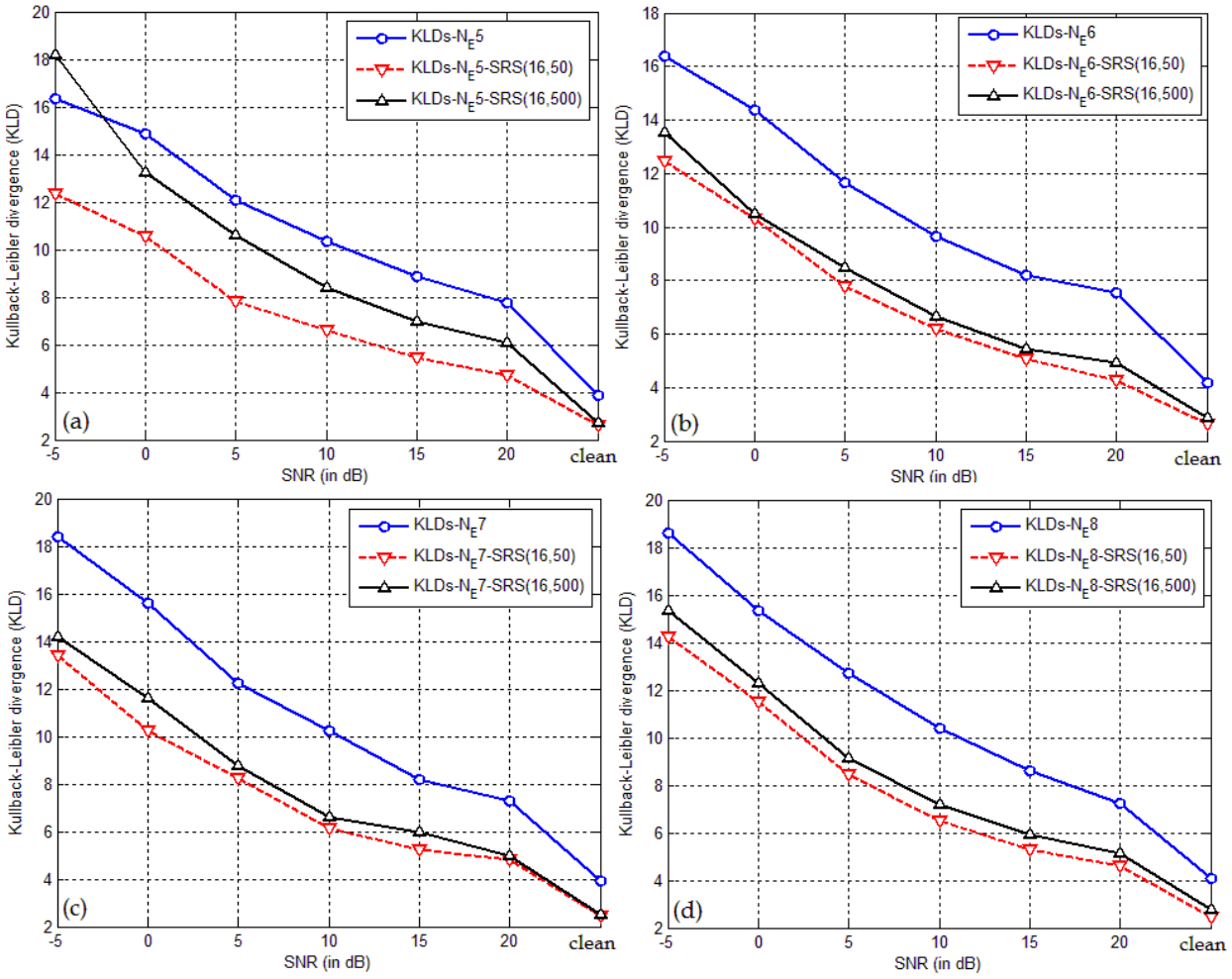

Figure 13: KLDs calculated between the PDFs of the speech feature vectors extracted from the testing and training speech signals. Each PDF was estimated from all the observations extracted from all the speech signals in the corresponding testing set that was used in the second EP, where the speech enhancement component was used. In Fig. 13(a), (b), (c) and (d), KLDs-N$_{\mathrm{E}}i$, $i = 5, \ldots, 8$ denote the KLDs calculated between the PDFs of speech feature vectors extracted from testing speech, contaminated by restaurant (N5), street (N6), airport (N7) and train station (N8) noises, respectively, and that of the speech feature vectors extracted from clean training speech. Similarly, the KLDs-N$_{\mathrm{E}}i$-SRS(16,50) and KLDs-N$_{\mathrm{E}}i$-SRS(16,500), $i = 5, \ldots, 8$, denote the KLDs calculated between the PDFs of speech feature vectors in the testing speech, which is SRS synthesized from enhanced speech signals, and the corresponding SRS for training.