Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation

Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, Atsunori Ogawa,
Takaaki Hori, Shinji Watanabe, Masakiyo Fujimoto, Takuya Yoshioka,
Takanobu Oba, Yotaro Kubo, Mehrez Souden, Seong-Jun Hahm, Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation, Japan
{marc.delcroix,kinoshita.k,nakatani.tomohiro,araki.shoko,ogawa.atsunori}@lab.ntt.co.jp

Abstract

In this paper, we introduce a system for recognizing speech in the presence of multiple rapidly time-varying noise sources. The main components of the proposed approach are a model-based speech enhancement pre-processor and an adaptation technique to optimize the integration between the pre-processor and the recognizer. The speech enhancement pre-processor consists of two complementary elements, a multi-channel speech-noise separation method that exploits spatial and spectral information, followed by single channel enhancement that uses the long-term temporal characteristics of speech. To compensate for any mismatch that may remain between the enhanced features and the acoustic model, we employ an adaptation technique that combines conventional MLLR with the dynamic adaptive compensation of the variance of the Gaussians of the acoustic model. Our proposed system greatly improves the audible quality of speech and substantially improves of the keyword recognition accuracy.

Index Terms: Robust ASR, Source separation, Model-based speech enhancement, Example-based enhancement, Model adaptation, Dynamic variance adaptation

1. Introduction

The problem of recognizing speech in the presence of multiple highly non-stationary noise sources remains a critical problem. Conventional approaches to noise robust speech recognition consist mainly of acoustic model compensation or speech/feature enhancement. Acoustic model compensation techniques are effective in mitigating the effect of stationary noise, but are difficult to employ in the presence of time-varying noise originating, for example, from interfering sources (TV, speaker...) or reverberation. On the other hand, many speech enhancement techniques have been developed to cope with non-stationary noise.

In this paper, we propose a system for recognizing speech in the presence of rapidly time-varying noise such as in the PASCAL 'CHiME' speech separation and recognition challenge [1]. To deal with these challenging noisy conditions, it is essential to use any information that may be available about the speech and noise. Most conventional speech enhancement or robust ASR systems use spatial [2, 3], spectral [4, 5] or temporal information [6, 7]. There have been only a few proposals that integrate several such information sources [8]. In this paper, we propose a system that fully employs spatial (locational), spectral and temporal information about the speech and noise, by using two complementary enhancement blocks.

First multi-channel speech-noise separation is performed using locational and spectral models of speech and noise [8, 9]. We adopt a method called dominance based locational and power-spectral characteristics integration (DOLPHIN). With DOLPHIN, rapidly changing speech and noise can be distinguished appropriately based mainly on their locational features, while the spectral shapes of the speech can be estimated reliably based mostly on the spectral features. With the CHiME challenge, since the target speaker location is fixed and speech and noise training data are available, the models can be trained in advance to achieve optimal performance.

Then long-term temporal information about the speech is used to further reduce non-stationary noise. This is achieved using an example-based enhancement algorithm [6, 7]. Example-based enhancement uses a parallel corpus containing speech sentences processed with the speech-noise separation algorithm and the corresponding clean speech. Enhancement is performed by searching for the longest speech segments in the corpus that best match the separated input speech. Then we use the corresponding clean speech segments to reconstruct the target speech with Wiener filtering. Using such long-term temporal information enables us to distinguish non-stationary noise from speech, thereby achieving high-quality enhancement.

The proposed speech enhancement system can greatly improve the quality and intelligibility of speech. With the CHiME challenge we are mostly interested in improving ASR performance. Therefore, we combine speech enhancement with state of the art techniques for recognition, such as the discriminative training of the acoustic model [10] and system combination [11]. Moreover, the interconnection of the pre-processor with the recognizer is achieved with the dynamic variance adaptation (DVA) technique to reduce any remaining mismatch between the enhanced speech and the acoustic model used for recognition [12]. DVA is similar to uncertainty decoding [13], in the sense that a dynamic (i.e. time-varying) feature variance is added to the acoustic model variance during decoding to mitigate the effect of unreliable features. It is based on a simple dynamic feature variance model that provides a general formulation enabling its use with many speech enhancement methods. Moreover, the model parameters are optimized for recognition using an adaptation technique. Consequently, the proposed recognition system achieves high recognition performance even under severe noise conditions, i.e. more than 90 % average keyword accuracy.

2. System overview

Figure 1 is a schematic diagram of the proposed recognition system. It consists of the following modules,

- Speech-noise separation based on DOLPHIN (see section 3.1).

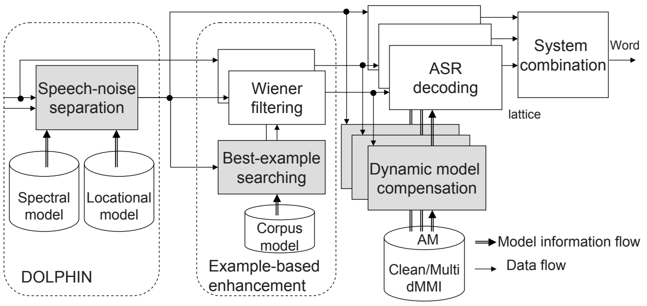

Figure 1: Proposed recognition system.

- Example-based speech enhancement (see section 3.2). Enhancement is applied to the observed noisy speech (Example-based I) and the speech processed with DOLPHIN (Example-based II), thus generating two enhanced speech signals.
- Dynamic acoustic model compensation based on DVA (see section 4).
- Speech recognizer using clean and multi-condition acoustic models (AM) trained using the differenced maximum mutual information (dMMI) discriminative criterion [10]. dMMI is a generalization of the minimum phone error (MPE) criterion that achieves superior or equivalent performance while being simpler to implement. Recognition is performed in parallel using speech processed with DOLPHIN, and the two enhanced outputs of the example-based speech enhancement.
- System combination to combine the different recognition outputs [11]. Each speech enhancement output is separately processed by the recognizer to output lattice results. These lattices are then combined by taking account of the word posterior probabilities.

In the following sections, we describe the enhancement and model compensation modules in more detail.

3. Speech enhancement

3.1. Speech-noise separation using DOLPHIN

To cope well with highly non-stationary noise such as that occurring in the CHiME challenge, DOLPHIN introduces statistical models of locational and spectral characteristics of both speech and noise[1]. All the models are assumed to be trained in advance using the CHiME challenge training data set, and utilized in a unified manner for speech-noise separation.

Suppose $X_{j,k}$ is a short time Fourier transform of a signal captured at a microphone j $(= 1, 2)$ and at a frequency k $(= 1, \ldots, N_k)$. Note that the time frame indices of all symbols are omitted in section 3.1 for the sake of notation simplicity. Then, the observed signal can be modeled as

$$X_{j,k} = \sum_l S_{j,k}^{(l)}, \tag{1}$$

where $S_{j,k}^{(l)}$ for $l = 1$ and $l = 2$, respectively, are the speech and noise signals captured at the j-th microphone. In this section, l is used as the index of the two sources, namely the speech $(l = 1)$ and the noise $(l = 2)$.

DOLPHIN uses two types of observed features: one is level normalized 2-ch observed signals, denoted as $\mathbf{d}_k$, and the other is the log power spectra of 1-ch signals obtained by applying delay-and-sum beamforming to the 2-ch observed signals to enhance the front signal, denoted as x_k. Letting T and $|\cdot|$, respectively, be the non-conjugate transpose of a vector and the Euclidean norm of a vector, the two features are defined as

$$\begin{aligned} \mathbf{d}_k &= \mathbf{X}_k/|\mathbf{X}_k|, \text{ where } \mathbf{X}_k = [X_{1,k}, X_{2,k}]^T \\ x_k &= \ln(|\sum_j X_{j,k}|^2). \end{aligned}$$

Because $\mathbf{d}_k$ represents the difference between channels, including interchannel phase and level differences, it is referred to as a location feature.

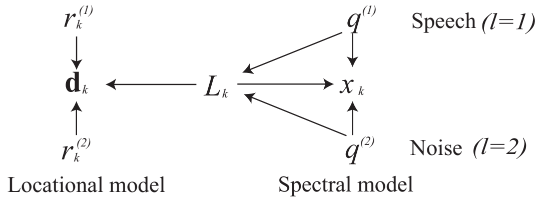

Figure 2: Graphical model of DOLPHIN.

DOLPHIN estimates the target speech $s_k^{(1)} = \ln(|\sum_j S_{j,k}^{(1)}|^2)$ for all k values, based on the above features. For this purpose, DOLPHIN introduces a generative model of the observed features as illustrated in Fig. 2. The model is composed of two sub-models shown in the left and right of the figure, which correspond to two generative models for $\mathbf{d}_k$ and x_k, respectively. To integrate the two sub-models, DOLPHIN utilizes a dominant source index (DSI) L_k that indicates whether speech or noise is more dominant at each frequency k. By sharing the DSI between the two sub-models, we can estimate the DSIs more reliably, and thus estimate the parameters of the sub-models more appropriately. In the following, we briefly describe the two sub-models and the parameter estimation method for the integrated generative model.

3.1.1. Sub-model for spectral feature

First, DOLPHIN models the log power spectra of speech and noise, denoted by $s_k^{(l)}$, by using spectral Gaussian mixture models (GMM). With a spectral GMM, the distribution of $s_k^{(l)}$ for each source l is modeled as $p(s_k^{(l)}; \psi_k^{(l)}) = \sum_q u_q^{(l)} \beta_k^{(l)}(s_k^{(l)}, q)$, where $\beta_k^{(l)}(s, q) = p(s|q; \psi_k^{(l)})$ is a Gaussian component indexed by q with a model parameter set $\psi_k^{(l)}$, and $u_q^{(l)}$ is its mixture weight. $\psi_k^{(l)}$ is assumed to be fixed in advance by prior training using the training data set.

To model the relationship between the source signal $s_k^{(l)}$ and the observed signal x_k, we adopt the log-max model [14] because it allows us to achieve efficient optimization based on the EM algorithm as discussed in [8]. The relationship is defined as $x_k = \max_l s_k^{(l)}$. Then, given the spectral Gaussian index pair, $\mathbf{q} = [q^{(1)}, q^{(2)}]$, the joint probability density function (pdf) of x_k and L_k is derived as

$$p(x_k, L_k = l|\mathbf{q}) \;=\; \beta_k^{(l)}(x_k, q^{(l)}) \int_{-\infty}^{x_k} \beta_k^{(l')}(s, q^{(l')}) ds,$$

where l' indicates the non-dominant source index.

3.1.2. Sub-model for locational feature

Let $\mathbf{D}_k^{(l)} = \mathbf{S}_k^{(l)}/|\mathbf{S}_k^{(l)}|$ be a location feature for a source l at a frequency k, where $\mathbf{S}_k^{(l)} = [S_{1,k}^{(l)}, S_{2,k}^{(l)}]^T$. According to the

[1] See [9] for more details about DOLPHIN.

sparseness assumption [2], we assume that the observed location feature $\mathbf{d}_k$ is equal to $\mathbf{D}_k^{(l)}$ of the dominant source at each frequency. Then, the posterior pdf of $\mathbf{d}_k$ given L_k can be rewritten as

$$p(\mathbf{d}_k|L_k = l) = p(\mathbf{D}_k^{(l)} = \mathbf{d}_k; \phi_k^{(l)}), \tag{2}$$

where $p(\mathbf{D}_k^{(l)}; \phi_k^{(l)})$ is the pdf of $\mathbf{D}_k^{(l)}$, and $\phi_k^{(l)}$ is its model parameter. To model $\mathbf{d}_k$, we need to define $p(\mathbf{D}_k^{(l)}; \phi_k^{(l)})$.

For a point source, a model of the location feature, referred to as a location vector model (LM), is proposed in [3], and used for source separation. However, the pdf of $\mathbf{D}_k^{(l)}$ in the assumed scenario is more complex than that for source separation, because the probabilistic uncertainty of $\mathbf{D}_k^{(l)}$ is derived not only from the reverberation effect but also from the change of the noise source locations. To model the pdf of such complex location features, we use a location vector mixture model (LMM). The LMM for $\mathbf{D}_k^{(l)}$ of each source l at each frequency k is defined as $p(\mathbf{D}_k^{(l)}; \phi_k^{(l)}) = \sum_r w_{r,k}^{(l)} \gamma_{r,k}^{(l)}(\mathbf{D}_k^{(l)})$ where $\gamma_{r,k}^{(l)}(\mathbf{D}) = p(\mathbf{D}|r_k^{(l)}; \phi_k^{(l)})$ is an LM indexed by $r_k^{(l)}$ at a frequency k, and $w_{r,k}^{(l)}$ is its mixture weight [3]. For computationally efficient optimization based on the EM algorithm, we further assume that $\mathbf{D}_k^{(l)}$ and $\mathbf{D}_{k'}^{(l)}$ are statistically independent when $k \neq k'$.

This paper assumes that $\phi_k^{(l)}$ is trained on the training data set in advance, and also adapted to each observed noisy utterance. The prior training and the adaptation were accomplished based on the learning algorithm for LMMs given in [3] and on its extension with the incremental EM algorithm, respectively. The details will be discussed in a future study.

3.1.3. Model parameter estimation

DOLPHIN considers the DSIs, L_k, to be hidden variables and estimates the spectral Gaussian index pair $\mathbf{q}$ at each time frame by maximizing the likelihood function defined as,

$$\begin{aligned} \mathcal{L}(\mathbf{q}) &= \sum_{\{L_k\}} p(\{\mathbf{d}_k\}, \{x_k\}, \{L_k\}, \mathbf{q}) \\ &= \sum_{\{L_k\}} \left(\prod_k \left(p(\mathbf{d}_k|L_k) p(x_k, L_k|\mathbf{q}) \right) \prod_l u_{q^{(l)}}^{(l)} \right), \end{aligned}$$

where $\{\cdot\}$ represents a set of variables at all frequencies. Then, as in [8], the combinations of model parameters over different sources can be estimated by the EM algorithm in a computationally efficient manner. Finally, the speech-noise separation can be achieved based on a minimum mean square error estimation, to output an estimate of the target clean speech. The overall processing flow can be found in [9], so we omit it in this paper because of the limited space.

In the experiments, we used a relatively long window for the feature extraction, that is a 100 ms Hann window with a 25 ms shift, to capture features for reverberant signals appropriately. For both speech and noise, we fixed the number of mixture components at 256 and 4 in the spectral and locational models, respectively. For speech, speaker dependent spectral GMMs were prepared, while a speaker independent LMM was prepared for the locational model. As regards noise, a pair of spectral and locational models were trained on all the noise data in the training data set.

3.2. Example-based enhancement

Even if DOLPHIN is powerful in terms of performing speech-noise separation it may not completely suppress non-stationary noise. Recently, an example-based approach to speech enhancement has been developed to handle highly non-stationary noise by exploiting the long-term temporal characteristics of speech [6, 7]. Here we extend the method for use as a post-processing technique for DOLPHIN. The method uses a parallel speech corpus created using stereo data composed of speech processed with DOLPHIN and the corresponding clean speech. The longest possible segments that match the input speech are extracted from the corpus to estimate the target speech. The use of such long segments provides long-term temporal dynamic information that enables us to differentiate speech from non-stationary noise. In the context of the CHiME challenge, the test utterances consist of commands with fixed grammar that can be well represented by a speech corpus. Therefore, for the CHiME recognition task, example-based enhancement seems particularly suited to suppressing the remaining non-stationary noise of DOLPHIN.

The method can be summarized as follows. A GMM is used to represent speech processed with DOLPHIN. The GMM is trained using noisy training data processed with DOLPHIN as,

$$\mathcal{G} \quad = \quad \sum_{m=1}^{M} w(m) \underbrace{N(\mathbf{y}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}_{g(\mathbf{y}|m)}, \tag{3}$$

where $\mathbf{y}$ is an MFCC feature vector of the output of DOLPHIN, $g(\mathbf{y}|m)$ is the m-th Gaussian component with the mean $\boldsymbol{\mu}_m$ and the covariance $\boldsymbol{\Sigma}_m$, and $w(m)$ is the corresponding weight. M is the number of mixture components. Then, a state sequence is associated to the training data as follows,

$$\mathcal{M} = \{\mathcal{G}, m_i \; i = 1, 2, \ldots, I\}, \tag{4}$$

where m_i is the index of a Gaussian component $g(\mathbf{y}|m_i)$ in $\mathcal{G}$ that produces the maximum likelihood for the i-th frame feature, $\mathbf{y}_i^{tr}$, of the training data set, and I is the total number of frames of the training data. Hereafter, we call $\mathcal{M}$ a corpus model. The clean speech used in the corpus is stored as amplitude spectra as follows,

$$\mathcal{A} = \{\mathbf{A}_i : i = 1, 2, \ldots, I\}, \tag{5}$$

where $\mathbf{A}_i$ is the amplitude spectrum of the clean speech associated with the i-th frame feature, $\mathbf{y}_i^{tr}$, of the training data set.

Enhancement is performed by first searching for the longest sequence of the corpus, $\mathcal{M}_{u:u+\tau_{\max}}^t$, that matches a sequence of the input processed speech $\mathbf{y}_{t:t+\tau}$ as,

$$\begin{aligned} \mathcal{M}_{u:u+\tau_{\max}}^t &= \underset{\tau, \mathcal{M}_{u:u+\tau}}{\arg\max}\; p(\mathcal{M}_{u:u+\tau}|\mathbf{y}_{t:t+\tau}), && (6) \\ &\approx \underset{\tau, \mathcal{M}_{u:u+\tau}}{\arg\max}\; \frac{p(\mathbf{y}_{t:t+\tau}|\mathcal{M}_{u:u+\tau})}{p(\mathbf{y}_{t:t+\tau})} && (7) \end{aligned}$$

where $\mathcal{M}_{u:u+\tau} = \{\mathcal{G}, m_i : \; i = u, u+1, \ldots, u+\tau\}$ represents the sequence of Gaussian components modeling consecutive frames from u to $u+\tau$ in the training data set, and $\mathbf{y}_{t:t+\tau}$ represents a segment taken from time frame t to $t+\tau$ of the processed speech $\mathbf{y}$, i.e. $\mathbf{y}_{t:t+\tau} = \{\mathbf{y}_t, \ldots, \mathbf{y}_{t+\tau}\}$. We assumed that the prior probability of the corpus segment $p(\mathcal{M}_{u:u+\tau})$ is constant for all segments. The likelihood of the processed speech given the segment $\mathcal{M}_{u:u+\tau}$ is given by

$$p(\mathbf{y}_{t:t+\tau}|\mathcal{M}_{u:u+\tau}) = \prod_{v=0}^{\tau} g(\mathbf{y}_{t+v}|m_{u+v}). \tag{8}$$

The longest segments are calculated for each time frame of the input processed speech. The estimate of the target speech at time frame t, $\hat{S}_t$, is obtained by averaging the amplitude spectra of all the segments that include the time frame t as follows,

$$\hat{\mathbf{S}}_t = \frac{\sum_v \mathbf{A}_{u+t-v} p(\mathcal{M}^v_{u:u+\tau_{\max}} | \mathbf{y}_{v:v+\tau_{\max}})}{\sum_v p(\mathcal{M}^v_{u:u+\tau_{\max}} | \mathbf{y}_{v:v+\tau_{\max}})} \quad (9)$$

where $\mathbf{A}_{u+t-v}$ is the amplitude spectrum associated with $\mathcal{M}^v_{u:u+\tau_{\max}}$ that corresponds to time frame t.

Finally, enhanced speech is obtained by Wiener filtering, using the estimated target speech given in eq. (9) as in [6]. Wiener filtering is applied to the noisy speech directly (Example-based I) or to the speech processed by DOLPHIN (Example-based II).

The above discussion assumes a single set of training data. One of the problems with the example-based enhancement method is that the searching process becomes computationally expensive when the corpus model becomes large. Indeed, we need to search for the best sequence among all the utterances used to create the corpus model. A single corpus model covering all speaker and noise conditions would be very large and so greatly increase the complexity of the search process. The CHiME recognition challenge allows for a speaker dependent recognition system. Therefore, we created a separate corpus model for each speaker. Moreover, to reduce the search cost even more, we utilize a separate corpus model for each SNR level. An input utterance is enhanced using the corpus model that best represents the utterance, i.e. the corpus model that provides the maximum likelihood of the utterance given the GMM model $\mathcal{G}$ as shown in eq. (3).

The parameters used in the experiments were as follows. The feature vector for the GMM $\mathcal{G}$ consisted of 60th order MFCCs with a log energy term. The number of mixture components M was 4096. The frame length was 20 ms, and frame shift is 10 ms. The total number of frames in the training data, I, ranges from 512127 to 755650 depending on the target speaker.

4. Interconnection of speech enhancement and recognizer using dynamic variance adaptation

Speech processed with an enhancement algorithm usually contains some artifacts that are detrimental to ASR. Such artifacts are time-varying (i.e. *dynamic*), and thus cannot be fully compensated for with conventional model compensation approaches such as MLLR [15]. Recently, there have been several proposals for increasing robustness by replacing the point estimates of the enhanced features by a distribution with a dynamic feature variance. Assuming that the acoustic model is represented by HMMs with a state density modeled by GMMs, the probability of the enhanced MFCC feature vector, $\hat{\mathbf{s}}_t$, given an acoustic model HMM state n, can then be expressed as [12],

$$p(\hat{\mathbf{s}}_t|n) = \sum_{m=1}^{M_a} \omega_{n,m} N(\hat{\mathbf{s}}_t; \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m} + \boldsymbol{\Sigma}_{b_t}), (10)$$

where m is the Gaussian mixture component index, M_a is the number of Gaussian mixtures, $\omega_{n,m}$ is the mixture weight, and $\boldsymbol{\mu}_{n,m}$ and $\boldsymbol{\Sigma}_{n,m}$ are the mean vector and covariance matrix, respectively. $\boldsymbol{\Sigma}_{b_t}$ is the *dynamic feature variance* that can be interpreted as a measure of feature uncertainty. We assume that it is diagonal with diagonal elements $\sigma^2_{b_t,i}$, where i is the feature dimension index. Without the additional $\boldsymbol{\Sigma}_{b_t}$, eq. (10) is equivalent to conventional ASR. The use of the dynamic feature variance makes it possible to mitigate the effect of unreliable features on the recognition results, since for these features the corresponding $\sigma^2_{b_t,i}$ values are large, which reduces the likelihood of all HMM state n. Note that eq.(10) is similar to uncertainty decoding as described in [13], but the estimation of the dynamic feature variance differs.

We have recently proposed the following model for the dynamic feature variance [12],

$$\hat{\sigma}^2_{b_t,i} = \alpha_i^2 \underbrace{(u_{t,i} - \hat{s}_{t,i})^2}_{\hat{b}^2_{t,i}}, \quad (11)$$

where $\hat{\sigma}^2_{b_t,i}$ is the estimated dynamic feature variance, α_i is the *pre-processor uncertainty weight*, $u_{t,i}$ and $\hat{s}_{t,i}$ are the observed and enhanced speech features, respectively, for time frame t and feature dimension index i. $\hat{b}^2_{t,i}$ provides a time-varying feature variance root. Features are considered unreliable when the pre-processor removes a lot of acoustic distortion. The pre-processor uncertainty weight α_i measures the reliability of the speech enhancement pre-processor. If the speech enhancement introduces many artifacts and the enhanced features are therefore unreliable, the pre-processor uncertainty weights will be large.

In [12] we proposed estimating α_i using adaptation to obtain optimal values as follows,

$$\boldsymbol{\alpha} = \arg\max_{\boldsymbol{\alpha}} \left(p(\{\hat{\mathbf{s}}_t\}|W, \boldsymbol{\alpha}) p(W) \right), \quad (12)$$

where $\{\hat{\mathbf{s}}_t\}$ is a sequence of enhanced speech feature vectors, W is the word sequence corresponding to the feature sequence $\{\hat{\mathbf{s}}_t\}$, $\boldsymbol{\alpha}$ is the set of model parameters to be optimized, i.e. $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_F)$ and F is the dimension of the feature vector. Eq. (12) can be solved using the EM algorithm or with a gradient descent optimization method. The DVA algorithm is described in detail in [12].

DVA focuses on variance compensation but it can be combined with conventional mean adaptation techniques such as MLLR [15] to further improve the interconnection between the speech enhancement pre-processor and the recognizer. There are several approaches that can be used to combine MLLR and DVA. Here we performed three iterations of MLLR recursively followed by three iterations of DVA and repeated the process 20 times.

In the experiments we used unsupervised adaptation to estimate $\boldsymbol{\alpha}$, i.e. the word sequence W was obtained from a first recognition pass performed without adaptation.

5. Experimental results

5.1. Experimental settings

We used the speech recognizer platform SOLON [16], which was developed at NTT Communication Science Laboratories. We generated two types of speaker dependent acoustic models, one using 'clean' speech (reverberant only) and one using multi-condition training data.

The clean speech model consisted of conventional left-to-right HMMs with a total of 254 states each modeled by a Gaussian Mixture consisting of seven Gaussians. We added a silent and short pause model to the original model provided by the CHiME challenge organizers. The original models trained with HTK were retrained with SOLON using the dMMI discriminative criterion [10].

We created multi-condition data by adding background noise samples to the reverberant training data. The amount of training data was 42 times the amount of clean training data (seven noise environments obtained from the background noise data provided by the CHiME challenge[17] by six SNR levels). The multi-condition noisy data were then processed with the DOLPHIN enhancement algorithm. The obtained multi-condition training data were used to train acoustic models. For the multi-condition model, we did not use silent and short pause models because it did not provide any significant recognition improvement. We used 20 Gaussians per HMM state to cover the variability of the multi-condition training data. The multi-condition acoustic models were also trained using the dMMI discriminative criterion [10].

We used also speaker dependent, SNR independent, unsupervised adaptation to further reduce any mismatch between the input features and the acoustic model. We used all the test data (from all SNR levels) from a given speaker to generate labels that were used for adaptation. The adaptation combined DVA and MLLR with a diagonal transformation matrix to adapt the mean parameters of the Gaussians. Hereafter we refer to this adaptation process as Adap.

We evaluated the results in terms of keyword recognition accuracy using the evaluation script provided by the CHiME challenge organizers [17].

5.2. Results for development test set

Table 1 shows the keyword recognition accuracy for the development set when using clean (systems I to VI) and multi-condition acoustic models (systems VII to XII). Systems I, II, VII and VIII provide baseline results obtained with noisy speech (without any enhancement) using clean and multi-condition training with maximum likelihood (ML) and dMMI criteria. The clean ML baseline (system I) performs better than the baseline provided by the organizers of the challenge because of the use of the silent model and because of SOLON's handling of the sparse training data provided better speaker dependent models[2] [17]. The systems trained using dMMI (systems II and VIII) provided improvement compared with the ML systems (systems I and VII), especially for the multi-condition model. Indeed, with multi-condition training, dMMI can take advantage of the large amount of data. Therefore, in the following, we report the results using only acoustic models trained with dMMI. Note that the upper bound keyword recognition accuracy obtained by recognizing 'clean' speech (reverberant speech with no noise) using the clean model trained with dMMI was 96.75%.

The first part of Table 1 (systems III and VI) shows the recognition results for the recognition systems when using DOLPHIN (system III) and DOLPHIN + example-based enhancement with Wiener filtering applied to the noisy speech (DOLPHIN + EX I, i.e. system IV) with clean acoustic models. DOLPHIN (system III) already provided an average keyword accuracy improvement of more than 13%. Combining DOLPHIN with example-based enhancement (system IV) provided an additional improvement of more than 3%[3]. Using adaptation (MLLR combined with DVA) as shown with system V and VI, we further improved the keyword accuracy by 3% for DOLPHIN and 0.8% for DOLPHIN combined with example based enhancement. Note that for DOLPHIN, we confirmed that MLLR and DVA speparately achieved comparable performance improvements of around 2%, and combining them achieved an additional 1% improvement.

The second part of the Table 1 (systems IX to XII) shows enhancement results when using acoustic model trained with multi-condition training data. The multi-condition training data were obtained by processing the multi-condition noisy training data with the DOLPHIN algorithm. Using DOLPHIN with the multi-condition acoustic model, we obtained an average accuracy improvement of 4% compared with the multi-condition noisy baseline and 5% when using adaptation. We also investigated the use of multi-condition model with DOLPHIN + example-based enhancement (systems XIII and XIV). Here, we used the same acoustic model as for system VII, i.e. trained with training data processed with DOLPHIN and therefore we use example-based with Wiener filtering applied to the speech processed by DOLPHIN (DOLPHIN + Ex II). The multi-condition model does not match well with the speech processed with DOLPHIN + Example-based enhancement, therefore we observe a significant performance degradation if no adaptation is performed, but the performance can be recovered to some extent using adaptation (system XII). Note that we expect that performance would improve if we used a multi-condition model trained on the DOLPHIN+Example-based enhancement output, but due to the considerable complexity of example-based enhancement, we omitted this experiment.

In table 1 we highlight the best performance among systems I to XII using bold italics. We observe that systems XI achieved the best performance at almost all SNR levels. Even though the other systems perform worse than system XI, they may cause different types of errors and thus can be used as a different source of information to improve performance with system combination method [11]. The last part of table 1 shows results obtained with the system combination technique using the three different systems that provided the best performance, i.e. systems VI, XI and XII. For almost all SNR levels, the best performance (shown in bold in table 1) was obtained with system combination and an average absolute keyword improvement of up to 0.7% could be achieved.

Our approach does not only achieve a significant improvement in terms of recognition performance, but also provides a substantial noise reduction and increases the audible quality. We evaluated the improvement brought about by speech enhancement in terms of segmental SNR averaged over the six SNR conditions evaluated as in [9]. The average segmental SNR of the noisy speech was -1.6 dB. DOLPHIN improved the segmental SNR up to 5.6 dB and DOLPHIN + example-based enhancement improved the segmental SNR up to 5.8 dB. Enhancement was particularly effective at a low SNR. With -6 dB noise, the segmental SNR of the noisy speech was -5.6 dB, and it was improved to 3.8 and 4.3 dB with DOLPHIN and DOLPHIN + example-based enhancement, respectively. We provide also some sound samples at [18] that attest to the good speech enhancement performance.

5.3. Results for the evaluation test set

Table 2 shows the recognition results for the evaluation test set. For conciseness, we only provide the most relevant results. Note that even though all parameters setting was performed using the development set, we obtained a slightly better performance with the evaluation test set. These results confirm the robustness of the proposed recognition system.

6. Conclusions

In this paper we presented a system for speech recognition in environments with highly non-stationary noise. We showed that the proposed system could greatly improve the audible quality of speech and provide a great improvement in recognition performance. The proposed system was developed for the CHiME Challenge command recognition task, but it could be extended for use under broader conditions. In this case, one issue will be to relax the hypothesis used by the DOLPHIN speech-noise

[2] This was corroborated by observing that we obtained a baseline comparable to the challenge baseline using a speaker independent model but a more than 8% absolute accuracy improvement for the speaker dependent acoustic models trained using SOLON.

[3] Note that the example-based algorithm uses multi-condition data and therefore strictly speaking the whole system does not rely only on clean training data.

Table 1: *Keyword recognition accuracy in percent for the development test set. The ‘clean’ baseline achieved 96.75% keyword accuracy.*

System	Model	Speech	Adap.	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
I	ML-clean	noisy	-	49.75	52.58	64.25	75.08	84.25	90.58	69.42
II	dMMI-clean	noisy	-	50.42	53.58	63.33	75.25	84.58	90.50	69.61
III	dMMI-clean	DOLPHIN	-	71.33	76.92	82.08	87.42	90.92	91.75	83.40
IV	dMMI-clean	DOLPHIN + EX I	-	77.42	80.92	84.17	89.42	92.33	***94.50***	86.46
V	dMMI-clean	DOLPHIN	X	77.08	81.42	86.83	89.33	92.42	93.42	86.75
VI	dMMI-clean	DOLPHIN + EX I	X	78.58	81.83	85.50	90.58	92.83	94.33	87.28
VII	ML-multi	noisy	-	69.75	75.08	83.25	86.33	92.00	92.75	83.19
VIII	dMMI-multi	noisy	-	73.25	78.08	84.92	87.75	92.08	93.67	84.96
IX	dMMI-multi	DOLPHIN	-	82.75	85.42	89.17	91.25	92.00	92.67	88.88
X	dMMI-multi	DOLPHIN + EX II	-	75.00	77.67	80.92	87.17	89.00	89.75	83.25
XI	dMMI-multi	DOLPHIN	X	***83.83***	***87.33***	***90.25***	***91.50***	***93.83***	93.75	***90.08***
XII	dMMI-multi	DOLPHIN + EX II	X	82.33	86.50	88.50	91.33	***93.83***	93.33	89.30
System Combination (VI + XI + XII)				**84.33**	**88.58**	90.17	**92.33**	**94.50**	**95.00**	**90.82**

Table 2: *Keyword recognition accuracy in percent for the evaluation test set.*

System	Model	Speech	Adap.	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Mean
II	dMMI-clean	noisy	-	45.67	52.67	65.25	75.42	83.33	91.67	69.00
III	dMMI-clean	DOLPHIN	-	71.58	77.92	85.08	90.25	91.58	93.92	85.06
IV	dMMI-clean	DOLPHIN + EX I	-	79.83	82.25	89.75	91.92	92.42	94.92	88.52
V	dMMI-clean	DOLPHIN	X	78.33	82.50	87.42	91.67	93.17	94.83	87.99
VI	dMMI-clean	DOLPHIN + EX I	X	80.42	82.58	90.00	92.58	92.75	***95.00***	88.89
VIII	dMMI-multi	noisy	-	70.58	77.75	84.92	89.42	91.50	94.00	84.70
IX	dMMI-multi	DOLPHIN	-	84.25	86.17	90.92	92.58	93.67	93.75	90.22
X	dMMI-multi	DOLPHIN + EX II	-	76.50	79.33	85.00	87.50	89.58	89.67	84.60
XI	dMMI-multi	DOLPHIN	X	**85.83**	***87.92***	***91.17***	***93.58***	***94.08***	94.17	***91.13***
XII	dMMI-multi	DOLPHIN + EX II	X	83.58	87.00	90.33	92.33	93.25	93.92	90.07
System Combination (VI + XI + XII)				85.58	**88.33**	**92.33**	**93.67**	**94.17**	**95.83**	**91.65**

separation method as regards the known location of the target source location. Another issue is to confirm whether the example-based algorithm can provide equivalently good performance with more complex tasks such as spontaneous speech, when the corpus utterances may not fully represent the test utterances.

7. References

[1] Christensen, H., Barker, J., Ma, N., and Green, P., "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," Proc. Interspeech'10, pp. 1918-1921, 2010.

[2] Yilmaz, O. and Rickard, S., "Blind separation of speech mixture via time-frequency masking," IEEE Trans. SP, vol. 52, no. 7, pp. 1830-1847, 2004.

[3] Sawada, H., Araki, S. and Makino, S., "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," Proc. WASPAA-2007, pp. 139-142, 2007.

[4] Moreno, P.J., Raj, B. and Stern, R.M., "A vector Taylor series approach for environment-independent speech recognition," Proc. ICASSP-96, vol. 2, pp. 733-736, 1996.

[5] Kristjansson, T., Hershey, J., Olsen, P. and Gopinath, R., "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," Proc. ICSLP'06, pp. 97-100, 2006.

[6] Ming, J., Srinivasan, R. and Crookes, D., "A corpus-based approach to speech enhancement from nonstationary noise," IEEE Trans. ASLP, vol. 19, no. 4, pp. 822-836, 2011.

[7] Kinoshita, K., Souden, M., Delcroix, M. and Nakatani, T., "Single channel dereverberation using example-based speech enhancement with uncertainty decoding technique," to appear in Proc. Interspeech-2011, 2011.

[8] Nakatani, T., Araki, S., Yoshioka, T. and Fujimoto, M., "Joint unsupervised learning of hidden Markov source models and source location models for multichannel source separation," Proc. ICASSP'11, pp. 237-240, 2011.

[9] Nakatani, T., Araki, S., Delcroix, M., Yoshioka, T. and Fujimoto, M., "Reduction of highly nonstationary ambient noise based on spectral and locational characteristics of speech and noise for robust ASR," to appear in Proc. Interspeech-2011, 2011.

[10] McDermott, E., Watanabe, S. and Nakamura, A., "Discriminative training based on an integrated view of MPE and MMI in margin and error space," Proc. ICASSP'10, pp. 4894-4897, 2010.

[11] Evermann, G. and Woodland, P. C., "Posterior probability decoding, confidence estimation and system combination," Proc. NIST Speech Transcription Workshop, 2000.

[12] Delcroix, M., Nakatani, T. and Watanabe, S., "Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing," IEEE Trans. ASLP, vol. 17, no. 2, pp. 324-334, 2009.

[13] Droppo, J., Acero, A. and Deng, L., "Uncertainty decoding with SPLICE for noise robust speech recognition," Proc. ICASSP'02, vol. 1, pp. 57-60, 2002.

[14] Roweis, S.T., "Factorial models and refiltering for speech separation and denoising," Proc. EUROSPEECH-2003, pp. 1009-1012, 2003.

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

[16] Hori, T., Hori, C., Minami, Y. and Nakamura, A., " Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. ASLP, vol. 15, no. 4, pp. 1352-1365, 2007.

[17] "The PASCAL 'CHiME' Speech Separation and Recognition Challenge," `http://www.dcs.shef.ac.uk/spandh/chime/challenge.html`, Cited 17 February 2011.

[18] `http://www.kecl.ntt.co.jp/icl/signal/kinoshita/publications/CHiME_demo/index.html`