# SYNTHESIZED STEREO-BASED STOCHASTIC MAPPING WITH DATA SELECTION FOR ROBUST SPEECH RECOGNITION

*Jun Du, Qiang Huo*

Microsoft Research Asia, Beijing, P. R. China
{jundu,qianghuo}@microsoft.com

## ABSTRACT

In this paper, we present a synthesized stereo-based stochastic mapping approach for robust speech recognition. We extend the traditional stereo-based stochastic mapping (SSM) in two main aspects. First, the constraint of stereo-data, which is not practical in real applications, is relaxed by using HMM-based speech synthesis. Then we make feature mapping more focused on those incorrectly recognized samples via a data selection strategy. Experimental results on Aurora3 databases show that our approach can achieve consistently significant improvements of recognition performance in the well-matched (WM) condition among four different European languages.

**Index Terms**: stereo-based stochastic mapping, HMM-based speech synthesis, data selection.

## 1. INTRODUCTION

With the progress of automatic speech recognition (ASR), the noise robustness of speech recognizers attracts more and more attentions for practical recognition systems. Many techniques [11] have been proposed to handle the difficult problem of mismatch between training and application conditions. One type of approaches to dealing with the above problem is the so-called feature compensation approach by using stereo data to learn the mapping function between clean speech and noisy speech. SPLICE [8], namely stereo-based piecewise linear compensation for environments, is one successful showcase which is an extension of techniques [1, 12] developed at Carnegie Mellon University (CMU) in the past decades. Recently, a stereo-based stochastic mapping (SSM) technique[2] is proposed, which outperforms SPLICE. The basic idea is to build a GMM for the joint distribution of the clean and noisy speech by using stereo data. The simplicity to construct a joint GMM without environment selection makes SSM easier to implement in recognition stage.

One main problem of these approaches is the constraint of stereo data. Several works are presented to address this issue. In [15], stochastic vector mapping (SVM), which represents the mapping from the noisy speech to clean speech by a simple transformation, is a generalized definition of SPLICE. And a joint training of the parameters of SVM function and HMMs is implemented by adopting maximum likelihood (ML) or minimum classification error (MCE) criteria. MMI-SPLICE [9] is much like SPLICE, but without the need for target clean features. Instead of learning a speech enhancement function, it learns to increase recognition accuracy directly with a maximum mutual information (MMI) objective function. FMPE [13], a kind of discriminatively trained features, is related with SPLICE to a certain extent [7].

The motivation of our approach is to relax the constraint of recorded stereo-data from a new viewpoint: synthesized pseudo-clean features generated by exploiting HMM-based synthesis method [14, 16] is used to replace the ideal clean features from one of the stereo channels in those stereo-based approaches. In [10], we demonstrate this approach can achieve even better performance than SPLICE in the clean training condition of Aurora2 database. In this work, we apply the synthesized features to SSM approach, and verify its effectiveness over a high-performance baseline of real-world ASR, namely the well-matched condition of Aurora3 databases. Actually, [2] has already shown that the performance gain in the multi-training condition is not significant. In our experiment, similar observations are made if the synthesis method in [10] is directly applied. To achieve better performance, a simple data selection strategy is designed to make the feature mapping more focused on those incorrectly recognized samples by comparing two synthesized feature sequences using correct labels and recognition results, respectively. Then our synthesized stereo-based stochastic mapping (SSSM) with data selection can achieve consistently significant improvements of recognition performance in the well-matched (WM) condition among four languages on Aurora3 databases.

The remainder of the paper is organized as follows. First we give a review of SSM in Section 2. In Section 3, we propose our synthesized stereo-based stochastic mapping approach with data selection. In Section 4, we report experimental results. Finally we conclude the paper in Section 5.

## 2. REVIEW OF SSM

Assume we have a set of stereo data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, where $\boldsymbol{x}$ is the clean feature representation of speech, and $\boldsymbol{y}$ is the corresponding noisy feature representation. $D$ is the dimension of feature vectors. Define $\boldsymbol{z} \equiv (\boldsymbol{x}, \boldsymbol{y})$ as the concatenation of the two channels. In the most general case, $\boldsymbol{y}$ representing $L_n$ noisy speech vectors is used to predict $\boldsymbol{x}$ representing $L_c$ clean speech vectors. To construct the mapping function between $\boldsymbol{y}$ and $\boldsymbol{x}$, the joint distribution $p(\boldsymbol{z})$ should be trained. Here Gaussian mixture model (GMM) is used:

$$p(\boldsymbol{z}_t) \quad = \quad \sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{z,k}, \boldsymbol{\Sigma}_{zz,k}) \qquad (1)$$

where $K$ is the number of mixture components, $c_k$, $\boldsymbol{\mu}_{z,k}$, and $\boldsymbol{\Sigma}_{zz,k}$, are the mixture weights, means, and covariances of each component, respectively. Then the mean vector $\boldsymbol{\mu}_{z,k}$ will be of dimension $D(L_c + L_n)$ and the covariance matrix $\boldsymbol{\Sigma}_{zz,k}$ will be of size $D(L_c + L_n) \times D(L_c + L_n)$. Also the mean and covariance can be

978-1-4673-2507-3/12/$31.00 © 2012 IEEE

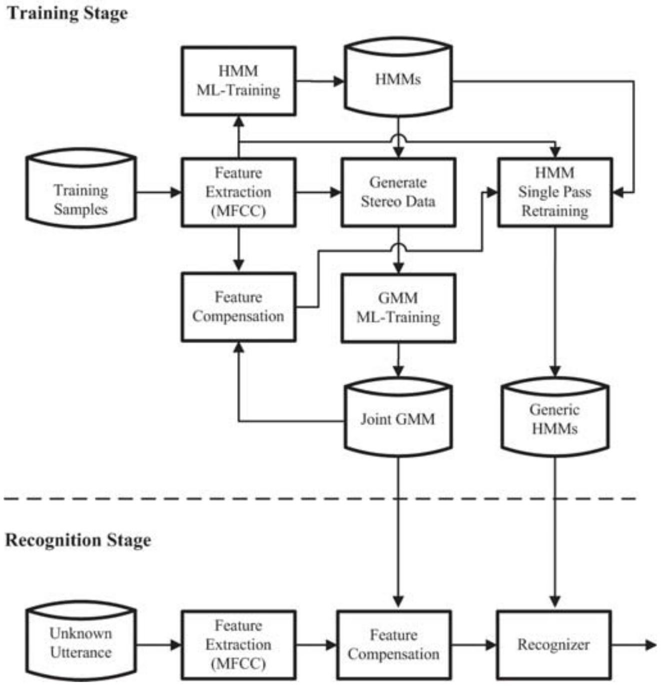

**Fig. 1**. Overall development flow and architecture.

partitioned as

$$\boldsymbol{\mu}_{z,k} = \begin{pmatrix} \boldsymbol{\mu}_{x,k} \\ \boldsymbol{\mu}_{y,k} \end{pmatrix} \tag{2}$$

$$\boldsymbol{\Sigma}_{zz,k} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx,k} & \boldsymbol{\Sigma}_{xy,k} \\ \boldsymbol{\Sigma}_{yx,k} & \boldsymbol{\Sigma}_{yy,k} \end{pmatrix} \tag{3}$$

where subscripts $x$ and $y$ indicate the clean and noisy speech respectively.

The above joint GMM distribution can be estimated in a classical way using EM algorithm. In the feature compensation stage, two estimation criteria, namely minimum mean-squared error (MMSE) and maximum a posteriori (MAP), can be applied. In this work, MMSE estimation is adopted:

$$\hat{\boldsymbol{x}} = E_x\left[\boldsymbol{x}|\boldsymbol{y}\right] = \sum_{k=1}^{K} P(k|\boldsymbol{y}) E_x\left[\boldsymbol{x}|\boldsymbol{y},k\right] \tag{4}$$

where $P(k|\boldsymbol{y})$ is the posterior probability defined as

$$P(k|\boldsymbol{y}) = \frac{c_k \mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_{y,k},\boldsymbol{\Sigma}_{yy,k})}{\sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_{y,k},\boldsymbol{\Sigma}_{yy,k})} \tag{5}$$

and the conditional expectation $E_x\left[\boldsymbol{x}|\boldsymbol{y},k\right]$ can be calculated as

$$E_x\left[\boldsymbol{x}|\boldsymbol{y},k\right] = \boldsymbol{\mu}_{x,k} + \boldsymbol{\Sigma}_{xy,k}\boldsymbol{\Sigma}_{yy,k}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{y,k}) \tag{6}$$

## 3. OUR APPROACH

The overall flowchart of our SSSM approach is illustrated in Fig. 1. In the training stage, first a baseline system can be trained using MFCC features with cepstral mean normalization (CMN). Then the stereo feature vectors are generated via the training features and baseline HMMs, which are used to train the joint GMM. Followed by feature compensation to training features using SSM, generic HMMs are generated by using single pass retraining (SPR) [17]. In the recognition stage, after feature compensation to MFCC features extracted from the unknown utterance, the normal recognition is performed. In the following sections, we elaborate on modifications to original SSM formulation and the generation of synthesized features.

### 3.1. SSM Modification

In SSM approach, an important step is to use Eq.(4) to estimate clean speech under MMSE criterion. In [2], it is indicated that the item $\boldsymbol{\Sigma}_{xy,k}\boldsymbol{\Sigma}_{yy,k}^{-1}$ in Eq.(6) represents the linear transformation to the noisy speech features. But according to the experiments of our SSSM approach, we observe that this linear transformation can even result in poor recognition performance. One possible explanation is although the covariance parameters $\boldsymbol{\Sigma}_{xy,k}$ and $\boldsymbol{\Sigma}_{yy,k}$ trained under the maximum likelihood criterion for feature compensation in Eq.(4) can bring the minimum squared error between clean and noisy speech features, it is not necessarily improve the discriminations among classes of the speech recognizer. So in our implementation of feature compensation, Eq.(6) is modified as

$$E_x\left[\boldsymbol{x}|\boldsymbol{y},k\right] = \boldsymbol{\mu}_{x,k} + (\boldsymbol{y}-\boldsymbol{\mu}_{y,k}) \tag{7}$$

which means only using bias compensation to noisy speech features is more stable than adding the linear transformation in this case. Another benefit from this modification is that we only need to train a joint GMM with diagonal covariances, which can significantly reduce the number of model parameters.

Acoustic context expansion by using several noisy feature vectors to predict the clean feature vector is another trick to improve the recognition performance [2], which increases the size of joint GMM. To achieve improvement of recognition performance but not increasing the size of joint GMM, we apply the following smoothing operation after feature compensation:

$$\hat{\boldsymbol{x}}_t^{\text{smooth}} = \frac{\sum_{\tau=-\Delta}^{\Delta}(\Delta+1-|\tau|)\hat{\boldsymbol{x}}_{t+\tau}}{\sum_{\tau=-\Delta}^{\Delta}(\Delta+1-|\tau|)} \tag{8}$$

where $\hat{\boldsymbol{x}}_t$ is the compensated feature vector of the $t^{\text{th}}$ frame, and $\Delta$ is the size for context expansion. It is interesting that this simple operation plays a similar role to the acoustic context expansion in original SSM based on our experiments.

### 3.2. Generation of Synthesized Features

Suppose that we only have noisy speech as the training data in real applications. Then HMMs trained using those noisy features are noise-robust to some extent. To synthesize the features as the “clean” channel of the stereo data, first state-level force-alignment of training features with true labels is performed. With this state sequence and corresponding HMMs, we can do the HMM-based speech synthesis [14]. The details of formulation can refer to [10]. Obviously, to the recognizer, those synthesized oracle feature sequences are perfectly matching and robust to not only noises, but also other irrelevant factors. A clearer illustration is given in Table 1, where the oracle features generated by HMM-based speech synthesis with true labels of both training and testing set are tested in the well-matched (WM) condition on Aurora3 databases. The word error rate in all cases are very low (most are less than 1%). This indicates that if a well-defined mapping function between the noisy feature and synthesized

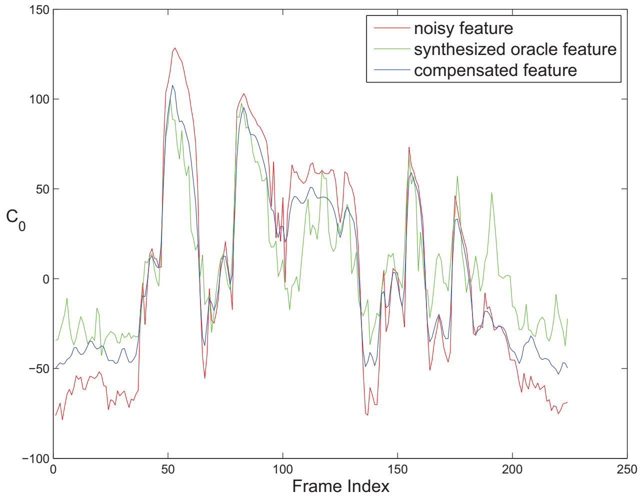

**Fig. 2**. Feature sequences of one training utterance from Aurora3 databases.

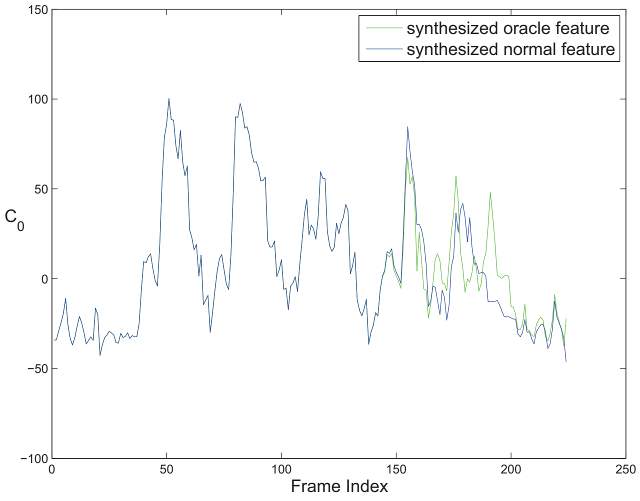

**Fig. 3**. Feature sequences of one training utterance from Aurora3 databases.

oracle feature can be given, the compensated system after feature mapping should achieve much higher recognition performance. Actually, this idea is partially verified in [10], where significant recognition performance gain is achieved in the clean-training condition. But in real applications, training set may consist of noisy speech data from multiple conditions, e.g., the well-matched (WM) condition of Aurora3 databases where a relative high recognition performance is already achieved for the testing set. Our preliminary experiments show that synthesized stereo stochastic mapping using oracle features can not yield consistently significant improvement of recognition performance among different languages. To get better insights, $C_0$ sequences of input noisy feature, synthesized oracle feature, and compensated feature generated from one representative training utterance of Aurora3 databases are compared in Fig. 2. In general, the envelope of compensated feature sequence is between the envelops of noisy feature and oracle feature, which means the compensated feature is truly approaching to oracle feature from noisy feature. If the "approaching" can not result in better recognition performance, one possible reason is that the uncertainty after compensation or modeling error is too large.

Inspired by those observations, we aim at improving our synthesized stereo-based stochastic mapping approach by data selection to make the feature mapping more focused on those incorrectly recognized samples. The location of region for incorrectly recognized samples can be illustrated in Fig 3, where synthesized oracle feature and synthesized normal feature are generated using HMM-based speech synthesis with true labels and recognition results, respectively. For the region where the envelops of synthesized oracle feature and synthesized normal feature are completely overlapped, the recognition results are correct. So the feature mapping should focus on the region where synthesized oracle feature and synthesized normal feature are different. The new synthesized feature $\boldsymbol{x}$ used as the "clean" channel of the stereo data can be generally formulated as:

$$\boldsymbol{x} = F(\boldsymbol{y}, \boldsymbol{x}^{\text{oracle}}, \boldsymbol{x}^{\text{normal}}) \tag{9}$$

where $\boldsymbol{y}$, $\boldsymbol{x}^{\text{oracle}}$, and $\boldsymbol{x}^{\text{normal}}$ denote the input noisy feature, synthesized oracle feature, and synthesized normal feature respectively. $F$ is a general function. In this work, a simple function is adopted:

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{x}^{\text{oracle}} - \boldsymbol{x}^{\text{normal}} \tag{10}$$

which represents an extreme case that "clean" channel is exactly the same as "noisy" channel for those correctly recognized regions while "clean" channel is calculated by adding the difference of synthesized oracle feature and synthesized normal feature to the "noisy" channel. For most ASR tasks with multiple condition training, the correctly recognized regions should be much more than the incorrectly recognized regions. To achieve maximum performance gain, stereo feature pairs for incorrectly recognized regions should be repeated several times to make them finally comparable to the correctly recognized regions.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Experimental Setup

In order to verify the effectiveness of the proposed approach on real-world ASR, Aurora3 databases are used, which contain utterances of digit strings recorded in real automobile environments for German, Danish, Finnish and Spanish, respectively. A full description of the above databases and the corresponding test frameworks are given in [3, 4, 5, 6].

In our ASR systems, each feature vector consists of 13 MFCCs (including $C_0$) plus their first and second order derivatives. The number of Mel-frequency filter banks is 23. MFCCs are computed based on power spectrum. CMN is applied to MFCC feature vectors. Each digit is modeled by a whole-word left-to-right CDHMM, which consists of 16 emitting states, each having 3 Gaussian mixture components. We focus on well-matched (WM) "training-testing" condition for experiments of Aurora3, where both training and testing data are recorded by close-talking (CT) and hands-free (HF) microphones. For stereo-based stochastic mapping, $K = 4096$, $D = 13$, $L_c = L_n = 1$. For acoustic context expansion, $\Delta$ is set to 1. In all the experiments, tools in HTK [17] are used for training and testing.

**Table 1**. Performance (word error rate in %) of oracle features generated by HMM-based speech synthesis with true labels in the well-matched (WM) condition on Aurora3 databases.

| | German | Danish | Finnish | Spanish |
|---|---|---|---|---|
| Training Set | 0.53 | 0.67 | 0.62 | 0.93 |
| Testing Set | 0.40 | 0.77 | 0.73 | 1.18 |

**Table 2**. Performance (word error rate in %) comparison of the baseline system and two feature compensation systems in the well-matched (WM) condition on Aurora3 databases.

| | German | Danish | Finnish | Spanish |
|---|---|---|---|---|
| Training Set | | | | |
| Baseline | 4.95 | 8.54 | 5.67 | 4.54 |
| SSSM | 4.46 | 8.45 | 5.24 | 4.43 |
| SSSM-DS | 2.76 | 6.65 | 3.01 | 3.14 |
| Testing Set | | | | |
| Baseline | 7.51 | 9.16 | 6.91 | 6.43 |
| SSSM | 6.81 | 8.53 | 6.65 | 6.70 |
| SSSM-DS | 6.59 | 8.00 | 5.44 | 5.87 |

### 4.2. Experimental Results

Table 2 summarizes a performance (word error rate in %) comparison of the baseline system and two feature compensation systems in the well-matched (WM) condition on Aurora3 databases. SSSM denotes the system where synthesized stereo-based stochastic mapping based on HMM-based speech synthesis in [10] is applied. SSSM-DS is the system where our proposed data selection strategy is used for SSSM. Several observations can be made. First, on the training set, our proposed SSSM-DS approach can achieve significant improvements of recognition performance compared with the baseline system, which is reasonable as the feature mapping focuses on those incorrectly recognized samples. Meanwhile the word error rate reduction for SSSM is marginal, which means that minimizing the error between the noisy feature and the oracle feature on the whole data set can not guarantee recognition performance boosting. Second, on the testing set, SSSM-DS consistently outperforms SSSM for all languages, especially Finnish and Spanish databases. Third, significant performance gain (overall more than 10% relative word error rate reduction) is achieved by SSSM-DS over the baseline system for all languages on the testing set. By comparing the performance gain of SSSM-DS on both training set and testing set, the generalization capability of SSSM-DS can be observed, which suggests that one possible future work should aim at further improving the recognition performance on the training set.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we investigate to make stereo-based stochastic mapping technique more practical for robust speech recognition. Synthesized features are generated to form the stereo data in SSM using HMM-based speech synthesis. A simple data selection strategy is adopted to make the feature mapping more focused on those incorrectly recognized samples. The effectiveness of the proposed approach has been confirmed in an experimental study on Aurora3 tasks. Ongoing and future works include 1) to study more theoretic formulation to the current intuitive data selection strategy, 2) to explore more advanced tool to model the nonlinear relationship between the stereo data, 3) to verify its effectiveness on large vocabulary speech recognition tasks, 4) to combine with other noise robust techniques.

## 6. REFERENCES

[1] A. Acero, *Acoustic and environment robustness in automatic speech recognition*, Kluwer Academic Publishers, 1993.

[2] M. Afify, X. Cui, and Y. Gao, “Stereo-based stochastic mapping for robust speech recognition,” *Proc. ICASSP*, 2007, pp.377-380.

[3] Aurora document AU/217/99, “Availability of Finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation,” Nokia, Nov. 1999.

[4] Aurora document AU/271/00, “Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: description and baseline results,” UPC, Nov. 2000.

[5] Aurora document AU/273/00, “Description and baseline results for the subset of the SpeechDat-Car German database used for ETSI STQ Aurora WI008 advanced DSR front-end evaluation,” Texas Instruments, Dec. 2001.

[6] Aurora document AU/378/01, “Danish SpeechDat-Car digits database for ETSI STQ-Aurora advanced DSR,” Aalborg University, Jan. 2001.

[7] L. Deng, J. Wu, J. Droppo, and A. Acero, “Analysis and comparison of two speech feature extraction/compensation algorithms,” *IEEE Signal Process. Lett.*, Vol. 12, No. 6, pp.477-480, 2005.

[8] J. Droppo, L. Deng, and A. Acero, “Evaluation of the SPLICE algorithm on the Aurora2 database,” *Proc. EuroSpeech*, 2001, pp.217-220.

[9] J. Droppo and A. Acero, “Maximum mutual information SPLICE transform for seen and unseen conditions”, *Proc. EuroSpeech*, 2005, pp.989-992.

[10] J. Du, Y. Hu, L.-R. Dai, and R.-H. Wang, “HMM-based pseudo-clean speech synthesis for SPLICE algorithm,” *Proc. ICASSP*, 2010, pp.4570-4573.

[11] Y. Gong, “Speech recognition in noisy environments: a survey,” *Speech Communication*, Vol. 16, No. 3, pp.261-291, 1995.

[12] P. J. Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[13] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, “fMPE: Discriminatively trained features for speech recognition,” *Proc. ICASSP*, 2005, pp.961-964.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. ICASSP*, 2000, pp.1315-1318.

[15] J. Wu and Q. Huo, “An environment-compensated minimum classification error training approach based on stochastic vector mapping,” *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 14, No. 6, pp.2147-2155, 2006.

[16] Z.-J. Yan, F. K. Soong, and R.-H. Wang, “Word graph based feature enhancement for noisy speech recognition,” *Proc. ICASSP*, 2007, pp. 373-376.

[17] S. Young *et al.*, The HTK Book (for HTK v3.4), 2006.