Missing-Feature Reconstruction With a Bounded Nonlinear State-Space Model

Ulpu Remes, Kalle J. Palomäki, Tapani Raiko, Antti Honkela, and Mikko Kurimo

***Abstract*—Missing-feature reconstruction can improve speech recognition performance in unknown noisy environments. In this work, we examine using a nonlinear state-space model (NSSM) for missing-feature reconstruction and propose estimation with observed bounds to improve the NSSM performance. Evaluated in large-vocabulary continuous speech recognition task with babble and impulsive noise, using observed bounds in NSSM state estimation significantly improved the method performance.**

***Index Terms*—Missing data, noise robustness, speech recognition, state space methods.**

I. Introduction

MISSING-FEATURE methods for automatic speech recognition are motivated by evidence on human speech perception in noise [1]. The methods assume a noisy speech signal can be divided in speech and noise dominated spectrotemporal components. The speech-dominated components are assumed reliable representations for the underlying clean speech signal whereas in the noise-dominated unreliable components, the clean-speech information is assumed lost. The missing-feature approach is rather well-suited for unpredictable noise conditions and sudden noise events, and while most work on the missing-feature methods has been conducted on limited-vocabulary data, reconstruction approaches such as cluster-based imputation [2] and sparse imputation [3] have proven effective in large-vocabulary continuous speech recognition as well [4].

In this work, we investigate missing-feature reconstruction based on nonlinear state-space modeling. Here, nonlinear dynamic factor analysis [5], [6] is applied for estimating a nonlinear state-space model (NSSM) for speech data. The missing values in the noisy speech signal are restored using the model and a state-sequence estimate calculated from the reliable speech components. The reconstruction performance was evaluated in [6] on clean-speech samples that contained 20–200 ms temporal gaps.

In noisy environments, where missing values coincide with the noise-dominated spectrotemporal components, the reliable values are few compared to the artificially constructed samples used in [6]. On the other hand, since real environmental noise is primarily additive, the clean-speech energy in the missing components can be assumed not to exceed the observed noisy-speech energy. To utilize this information in missing-feature reconstruction, we introduce the observed bounds into NSSM state estimation. The bounded NSSM method is evaluated in large-vocabulary continuous speech recognition task using i) clean speech samples with missing values distributed as described in [6] and ii) speech corrupted with either babble or impulsive noise. Babble noise is a common benchmark in noisy-speech recognition experiments and impulsive noise is assumed to be the best test bench for the temporal modeling capability of the NSSM method. NSSM has not been evaluated in a speech recognition task previously.

II. Methods

A. Missing-Feature Reconstruction

When speech signal is corrupted with additive noise from an uncorrelated source, the ith component of the τth noisy speech feature $Y(\tau, i)$ in log-compressed mel-spectral domain can be approximated as $Y(\tau, i) \approx \max\{X(\tau, i), N(\tau, i)\}$, where $X(\tau, i)$ denotes the clean speech and $N(\tau, i)$ the noise power. Missing-feature methods divide the noisy observations into reliable and unreliable components depending on whether the component is dominated by speech or noise. The reliable components $Y_r(\tau, i)$ in the noise-corrupted observations $\boldsymbol{Y}(\tau)$ are assumed fair estimates for the clean speech so that the corresponding clean speech values $X_r(\tau, i) \approx Y_r(\tau, i)$, whereas the clean-speech values $X_u(\tau, i)$ corresponding to the unreliable components are effectively missing. The unreliable components only provide an upper bound for the corresponding clean-speech values, $X_u(\tau, i) \leq Y_u(\tau, i)$. The so called reconstruction or imputation methods substitute the missing values with clean-speech estimates $\hat{X}_u(\tau, i)$ which are often calculated based on a model derived from a set of clean-speech training data.

B. Spectrographic Mask Estimation

Dividing the noisy speech signal into unreliable and reliable spectrotemporal components is referred to as spectrographic mask estimation. The approach used in this work is based on the negative energy criterion [1]. The observed features $Y(\tau, i)$ are considered reliable if $\exp(Y(\tau, i)/\hat{N}(\tau, i)) > \gamma$, where $\hat{N}(\tau, i)$ denotes a noise estimate and γ is a threshold parameter. For

Manuscript received June 15, 2011; accepted July 19, 2011. Date of publication August 04, 2011; date of current version August 22, 2011. This work was supported by Hecse, the Academy of Finland under the Centres of Excellence in Research Programme (129674) and Projects 114369, 121179, 121802, and 133145, and the IST Programme of European Community under the PASCAL2 Network of Excellence (IST-2007-216886). This publication only reflects the authors' views. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Murat Saraclar.

U. Remes, K. J. Palomäki, T. Raiko, and M. Kurimo are with the Adaptive Informatics Research Centre, Aalto University School of Science, Espoo, Finland (e-mail: ulpu.remes@tkk.fi).

A. Honkela is with the Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, Finland.

Digital Object Identifier 10.1109/LSP.2011.2163508

babble noise (see Section III-B for data set description), we use noise estimates calculated during speech pauses detected using a speech–nonspeech classifier as described in [4]. Here, the frames $\boldsymbol{Y}(\tau)$ classified as nonspeech are temporally smoothed to produce the noise estimate $\hat{\boldsymbol{N}}(\tau)$. For impulsive noise, we estimate the impulse locations in time domain and define the noise estimates as $\hat{\boldsymbol{N}}(\tau) = \boldsymbol{Y}(\tau)$ if impulse is detected in frame τ and $\hat{\boldsymbol{N}}(\tau) = \beta\hat{\boldsymbol{N}}(\tau-1)$ with $\beta = 0.985$ otherwise. Impulse detection is based on modeling speech as an autoregressive process and analyzing the model prediction error as proposed in [7]. The parameters used in mask estimation were optimized using the development data described in Section III-B.

C. Reconstruction With a Nonlinear State-Space Model

The clean-speech features in log-compressed mel-spectral domain are modelled as the output of a nonlinear state-space model (NSSM) where the observations $\boldsymbol{X}(\tau)$ are predicted from a hidden state variable $\boldsymbol{S}(\tau)$ as specified in the model equations

$$\boldsymbol{X}(\tau) = \boldsymbol{f}(\boldsymbol{S}(\tau)) + \boldsymbol{M}_1(\tau) \tag{1}$$

$$\boldsymbol{S}(\tau) = \boldsymbol{g}(\boldsymbol{S}(\tau-1)) + \boldsymbol{M}_2(\tau) \tag{2}$$

where the mappings $\boldsymbol{f}$ and $\boldsymbol{g}$ are nonlinear and $\boldsymbol{M}_1(\tau)$ and $\boldsymbol{M}_2(\tau)$ are Gaussian modeling error terms with zero mean and unknown diagonal covariance. In nonlinear dynamic factor analysis [5], the mappings are modelled using multilayer perceptron (MLP) networks and the model parameters estimated using variational Bayesian learning. Parameter estimation is based on minimizing a cost function that measures the Kullback–Leibler (KL) divergence between the approximate and true posteriors, which is equivalent to maximizing the lower bound of the log marginal likelihood $\log P(\boldsymbol{X})$. For a thorough discussion on the model and the variational learning approach, see [5].

Missing-feature reconstruction using NSSM was proposed in [6]. It is based on the learning approach described above, but the model parameters are fixed after training, and in the reconstruction phase, the KL cost function is evaluated over the noisy observations $\boldsymbol{Y}(\tau)$ and minimized with respect to the approximate posterior distributions of the state variable $\boldsymbol{S}(\tau)$. Starting from a random initialization, the evaluation and minimization steps are repeated in an iterative manner. The cost function is a sum of terms of which the most relevant to missing feature reconstruction is the prediction error term (cf. [5, eq. (5.7)]):

$$C \propto \sum_{\tau}\sum_{i} \sigma_M(\tau,i)^{-1}(\bar{X}(\tau,i) - Y(\tau,i))^2 \tag{3}$$

where $\bar{\boldsymbol{X}}(\tau) = \bar{\boldsymbol{f}}(\boldsymbol{S}(\tau))$ are the observation mean predictions at kth iteration. $\sigma_M(\tau,i) = \sigma_M(i)$ is the ith diagonal component in the covariance matrix learned for $\boldsymbol{M}_1(\tau)$ if $Y(\tau,i)$ is reliable and $\sigma_M(\tau,i) = \infty$ if $Y(\tau,i)$ is unreliable [6]. Thus, unreliable observations do not affect the prediction error. To improve state inference over long gaps of missing values, the partial derivatives of the cost function with respect to the state posteriors are replaced with approximate total derivatives; see [6] for details.

Final predicted observations $\bar{\boldsymbol{X}}^*$ are calculated based on the state sequence $\boldsymbol{S}^*$ that the NSSM estimation has converged to. The reconstructed features $\hat{X}(\tau,i) = Y(\tau,i)$ if $Y(\tau,i) \in \boldsymbol{Y}_r$ and $\hat{X}(\tau,i) = \bar{X}^*(\tau,i)$ if $Y(\tau,i) \in \boldsymbol{Y}_u$ as discussed in Section II-A. The resulting estimates $\hat{X}(\tau)$ are constrained not to exceed the observed upper bound as $\hat{X}(\tau,i) = \min\{\hat{X}(\tau,i), Y(\tau,i)\}$. In preliminary experiments, using the clean-speech estimates degraded speech recognition performance compared to using noisy observations unless the minimum rule was applied.

We have described so far the baseline NSSM method which does not use observed bounds in state estimation. In this work, we propose an approach for bounded NSSM state estimation. The observed bounds are introduced in state estimation by defining the component-specific variances in (3) as

$$\sigma_M(\tau,i) = \begin{cases} \sigma_M(i), & \text{if } Y \in \boldsymbol{Y}_r \text{ or } \bar{X} \geq Y \text{ or } \bar{X} \leq L \\ \infty, & \text{otherwise} \end{cases} \tag{4}$$

where $Y = Y(\tau,i)$ are the observed features and $\bar{X} = \bar{X}(\tau,i)$ the predicted clean-speech observations at kth iteration, and the lower bounds $L = L(i)$ are the estimated average log-mel-spectral energies in silence frames. Note that the variances $\sigma_M(\tau,i)$ are redetermined at every iteration, and the unreliable observations increase the prediction error (3) if their estimated values at current iteration lie outside the given bounds.

D. Other Missing-Feature Reconstruction Methods

Since NSSM has not been evaluated in a speech recognition task before, we compare the method performance to cluster-based imputation [2] and sparse imputation [3]. Cluster-based imputation represents clean speech frames as independent observations sampled from a Gaussian mixture model (GMM) whereas sparse imputation is a nonparametric method based on modeling clean speech segments as linear combinations of a limited number of example segments stored in a clean speech dictionary. The methods have previously performed well on the babble-noise data used in this work [4] and should thus provide competitive reference results for the NSSM approach. Sparse imputation also incorporates temporal modeling since features are processed in windows that span several time frames. Other methods that use time context in missing-feature reconstruction include e.g., correlation-based imputation [2] which is not used in this work since it has invariably resulted in lower performance rates than cluster-based imputation when evaluated on realistic data.

III. Experiments

A. System

The large-vocabulary continuous speech recognition system used in this work has been described in [8]. The speech signal is represented with 12 MFCC and a log energy feature and their first and second order differentials. The features are normalised with cepstral mean subtraction (CMS) and maximum likelihood linear transformation (MLLT). The acoustic models are state-clustered hidden Markov models for context-dependent triphones constructed using a phonetic decision tree. The output densities of the states are modelled as Gaussian mixtures and the durations by gamma distributions. The decoder is a time-synchronous beam-pruned Viterbi token-pass system and the language model a morph-based growing n-gram model trained on 145 million words of Finnish book and newspaper data.

Missing features are reconstructed in the 21-D log-compressed mel-spectral domain. The NSSM used in this work has a 7-D state space and the nonlinear mappings $\boldsymbol{f}$ and $\boldsymbol{g}$ in (4) and (5) are modelled as MLP networks with 30 and 20 neurons in the hidden layers, respectively. The parameters were estimated in 5000 training iterations and in reconstruction phase, the states were estimated in 200 iterations using the total derivatives approach proposed in [6]. With babble and impulsive noise data (see Section III-B), state estimation is initialized with five iterations during which the unreliable components are assumed reliable and their values fixed to the estimated average energy in silence, $L(i)$.

The GMM in cluster-based imputation has 10 Gaussian components with full covariance matrices, which results in a clean-speech model with approximately the same number of parameters as the NSSM. The model parameters are estimated using the expectation-maximization (EM) algorithm implemented in the GMMBAYES Matlab toolbox[1]. The bounded estimates are solved iteratively over the frequency channels as proposed in [2] with maximum of 200 iterations. The implementation and parameters for sparse imputation are as in [4] and the features are processed in 15-frame windows in both noise conditions. Finally, the spectrographic mask threshold is $\gamma = 4$ dB for the NSSM and cluster-based reconstruction methods and $\gamma = 5$ dB for the sparse imputation method in babble noise condition and $\gamma = 1$ dB for all methods in impulsive noise condition. The thresholds and the window width for sparse imputation were optimized using the development data described in Section III-B .

B. Data

The data used in this work is from the Finnish SPEECON database. Acoustic models are trained with a 30-h training set that contains clean speech recorded with a headset in quiet conditions. The NSSM and GMM models are trained with 500 read sentences (52 minutes) randomly selected from the SPEECON training set and the exemplar dictionary for sparse imputation sampled from 14 hours of read sentences. The speech–nonspeech classifier used in mask estimation is trained with babble-noise corrupted television news data from the Finnish Broadcasting Company (YLE).

In the previous work [6], the reconstruction performance of NSSM was evaluated on clean speech data with $T = 3$ and $T = 30$ consecutive frames missing in every 100 frames. In this work, we use clean speech samples corrupted with pink noise from NOISEX-92 with a signal-to-noise rate (SNR) 0 dB following the pattern used in [6]. The evaluation data corrupted with pink noise bursts consists of 350 sentences (36 minutes) from 39 speakers. To evaluate NSSM-based reconstruction in a realistic noisy-speech recognition task, additional evaluation sets are constructed from 1118 clean speech sentences (113 minutes) from 40 speakers. The utterances are artificially corrupted with babble noise from NOISEX-92 or with impulsive noise i.e., hammering recorded with a Sennheiser PC 130 headset 90 cm from the noise source (metal hammer on nail). The impulsive-noise data can be seen as a realistic counterpart to the pink noise burst data used in this work and the speech pause data used in [6]. The development data sets used in this work are constructed from 350 clean speech sentences (36 minutes) from 39 speakers. The utterances are corrupted with SNR 10 dB babble noise from NOISEX-92 or SNR 0 dB impulsive noise i.e., hammering recorded 30 cm from the noise source (metal hammer on wood).

TABLE I
RESULTS (LER/RMSE) ON PINK NOISE BURSTS (A) ASSUMING UNRELIABLE FRAMES AND (B) WHEN ORACLE MASKS ARE USED

		$T = 3$	$T = 30$
(a)	NSSM	3.9 / 0.04	33.5 / 0.09
	NSSM bounded	3.8 / 0.04	32.1 / 0.08
	CI	8.2 / 0.10	39.5 / 0.10
	SI	5.6 / 0.08	36.3 / 0.12
		$T = 3$	$T = 30$
(b)	NSSM	3.7 / 0.04	18.6 / 0.06
	NSSM bounded	3.7 / 0.04	11.8 / 0.06
	CI	4.0 / 0.05	13.0 / 0.05
	SI	4.1 / 0.07	10.1 / 0.06

Since words in Finnish are often long and consist of several morphemes, the speech recognition performance is measured in letter errors instead of word errors to have finer resolution for the results.

C. Results

Results from the pink noise burst experiments are presented in Table I. The missing features are estimated using NSSM, bounded NSSM, cluster-based imputation (CI), and sparse imputation (SI). The noise-corrupted frames are either (a) assumed completely unreliable or (b) oracle masks are calculated. The oracle masks assume that speech and noise signals are known a priori and label the spectrotemporal components reliable if the local SNR $\exp(X(\tau,i)/N(\tau,i)) > 0$ dB. The reconstruction methods are evaluated based on speech recognition performance measured in letter error rate (LER) and root mean squared error (RMSE) between the reconstructed and clean-speech features $\hat{\boldsymbol{X}}_u$ and $\boldsymbol{X}_u$ in the log-mel-spectral domain. Using bounded NSSM or NSSM results in the best speech recognition performance in most conditions except when $T = 30$ and oracle masks are used. Although bounded state estimation notably improves NSSM performance in this condition, the best results are obtained with sparse imputation. Moreover, in this condition, the lowest reconstruction error is obtained with cluster-based imputation. Note that as a frame-based method, it results in the same RMSE for $T = 3$ and $T = 30$.

Results from the experiments with babble and impulsive noise data are presented in Table II. The results are reported in LER and statistical significance in pairwise comparisons tested using the Wilcoxon signed rank test with each speaker-specific LER considered an observation. Pairwise comparisons are

[1] Available from www.it.lut.fi/project/gmmbayes

TABLE II
RESULTS (LER) ON (A) BABBLE AND (B) IMPULSIVE NOISE DATA

		SNR 15	SNR 10	SNR 5	SNR 0
	NSSM	10.8	34.8	73.9	85.4
(a)	NSSM bounded	8.2	20.7	54.3	76.9
	CI	8.3	21.6	54.9	78.5
	SI	7.1	16.4	43.3	72.3
		SNR 5	SNR 0	SNR -5	SNR -15
	NSSM	19.2	27.0	37.3	53.1
(b)	NSSM bounded	12.8	17.4	25.0	44.8
	CI	25.6	30.1	35.9	46.7
	SI	18.7	24.7	32.9	46.2

conducted on results from each noise condition separately. In all the experiments, bounded NSSM outperformed the baseline NSSM method with a statistically significant ($p < 0.0001$ in all comparisons) difference. Evaluated on the babble noise data, the differences between bounded NSSM and cluster-based imputation were not statistically significant ($(p = n.s.)$) at significance level $\alpha = 0.05$ in all except the SNR 0 dB condition, and the best results ($p < 0.0001$ in all pairwise comparisons) were obtained with sparse imputation. Evaluated on the impulsive-noise data, the best results ($p < 0.001$ in all pairwise comparisons) were obtained with bounded NSSM.

IV. DISCUSSION

Using a nonlinear state-space model (NSSM) for missing-feature reconstruction was proposed in [6]. In this work, we evaluated NSSM-based reconstruction in noise robust speech recognition task and proposed an approach for using the observed upper bounds to restrict the NSSM state estimation. Bounded state estimation significantly improved the NSSM performance under both babble and impulsive noise. The best results were obtained with sparse imputation in the babble-noise condition and with bounded NSSM in the impulsive-noise condition.

Although using either sparse imputation or bounded NSSM resulted in the best performance rates, the overall results suggest it is not temporal modeling that defines a method performance but differences in performance additionally stem from the difference between statistical and exemplar-based approaches. Evaluated on the babble-noise data, bounded NSSM and cluster-based imputation perform at the same level, but when the unreliable values are clustered in time, NSSM performance improves compared to cluster-based imputation. This is likely due to temporal modeling. However, although sparse imputation also employs temporal modeling, the difference in speech recognition performance between sparse imputation and cluster-based imputation narrows down in impulsive noise.

Evaluated on clean-speech data corrupted with pink-noise bursts, both NSSM methods resulted in speech recognition performance better than the reference methods when burst length $T = 3$. In this condition, the dynamic state estimation in NSSM can benefit from the reliable frames around the impulse-like noise bursts and the NSSM performance is almost same regardless to whether complete frames are assumed unreliable [Table I(a)] or oracle masks used [Table I(b)]. Sparse imputation performance, on the other hand, depends on using the oracle mask even in $T = 3$ case although the features are processed in multi-frame windows. When tested on babble or impulsive noise, the performance of sparse imputation relative to other methods improves with decreasing SNR, so the result is unlikely due to difficulties in dealing with too many missing values.

While for the most part, the spectrographic masks estimated for impulsive noise resemble the oracle masks calculated for $T = 30$ bursts in the pink noise burst scenario, they also contain one or more frames that are labelled completely unreliable. This is because the mask estimation method used in this work assumes that the impulse peak dominates over all clean speech information. Other noise types such as babble or pink noise typically do not mask the speech components with the highest energies. Since the distribution of high-energy components within a speech segment or window is likely a prominent cue for separating between different speech tokens, exemplar-based methods like sparse imputation may be sensitive to missing high-energy values. We believe this explains the differences between sparse imputation and NSSM that both employ temporal modeling but represent different modeling paradigms.

ACKNOWLEDGMENT

The authors would like to thank J. F. Gemmeke for the sparse imputation code.

REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, Jun. 2001.

[2] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, Sep. 2004.

[3] J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 272–287, Mar. 2010.

[4] J. F. Gemmeke, B. Cranen, and U. Remes, "Sparse imputation for large vocabulary noise robust ASR," *Comput. Speech Lang.*, vol. 25, no. 2, pp. 462–479, Apr. 2011.

[5] H. Valpola and J. Karhunen, "An unsupervised ensemble learning method for nonlinear dynamic state-space models," *Neural Comput.*, vol. 14, no. 11, pp. 2647–2692, Nov. 2002.

[6] T. Raiko, M. Tornio, A. Honkela, and J. Karhunen, "State inference in variational Bayesian nonlinear state-space models," in *Proc. ICA 2006*, Mar. 2006, pp. 222–229.

[7] S. V. Vaseghi and P. J. W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *Proc. Inst. Elect. Eng.*, vol. 137, no. 1, pp. 38–46, Feb. 1990.

[8] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Comput. Speech Lang.*, vol. 20, no. 4, pp. 515–541, Oct. 2006.