GMM-based missing-feature reconstruction on multi-frame windows

Ulpu Remes[1], *Yoshihiko Nankaku*[2], *Keiichi Tokuda*[2]

[1]Adaptive Informatics Research Centre, Aalto University School of Science, Finland
[2]Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Japan

Abstract

Methods for missing-feature reconstruction substitute noise-corrupted features with clean-speech estimates calculated based on reliable information found in the noisy speech signal. Gaussian mixture model (GMM) based reconstruction has conventionally focussed on reliable information present in a single frame. In this work, GMM-based reconstruction is applied on windows that span several time frames. Mixtures of factor analysers (MFA) are used to limit the number of model parameters needed to describe the feature distribution as window width increases. Using the window-based MFA in noisy speech recognition task resulted in relative error reductions up to 52 % compared to frame-based GMM.

Index Terms: factor analysis, missing feature methods, noise robustness, speech recognition

10.21437/Interspeech.2011-214

1. Introduction

Using missing-feature methods for automatic speech recognition in noisy conditions is based on finding reliable information in the noise corrupted speech signal. While marginalisation approaches [1] attempt to decode the noisy signal based on reliable observations alone, reconstruction or imputation methods substitute the noise-corrupted observations with clean-speech estimates prior to decoding. Reconstruction methods such as Gaussian mixture model (GMM) based reconstruction [2] and sparse imputation (SI) [3] have performed well in large-vocabulary continuous speech recognition (LVCSR) tasks under adverse noise conditions.

The GMM-based reconstruction method proposed in [2] processes each frame independently. The frame-based approach is sufficient in moderate noise conditions where most speech frames retain reliable information in some frequency channels, but limits the method performance in others. Experiments reported in [4] indicate that using GMM reconstruction results in significantly lower performance than using the window-based SI method when the signal-to-noise ratio (SNR) is low and few reliable features remain. Frame-based processing is also unsuited for speech corrupted with loud impulsive noise since impulses often mask complete time frames.

In this work, we investigate using GMM-based reconstruction with a moving window approach. Estimates for the missing values are calculated using all the reliable information within a fixed-length window spanning multiple frames. To limit the number of model parameters in the window-based model, we propose using a mixture of factor analysers (MFA) [5] in place of the normal GMM. Window-based GMM/MFA reconstruction is evaluated in LVCSR task using speech data corrupted with babble and impulsive noise at several SNRs. Performance with the SI method [3, 4] is reported for reference.

2. Methods

2.1. Mask estimation

Missing-feature methods divide the noisy observations into reliable and unreliable spectrotemporal components depending on whether the component is dominated by speech or noise. The approach used in this work is based on the negative energy criterion proposed in [1]. We denote the ith frequency channel in the τth noisy speech frame in the log-compressed mel-spectral domain as $Y(\tau, i)$ and the corresponding noise estimate as $\hat{N}(\tau, i)$. The observed features $Y(\tau, i)$ are considered reliable if $\exp(Y(\tau, i)/\hat{N}(\tau, i)) > \gamma$, where γ is a threshold parameter. The resulting label matrix that divides the spectrogram in reliable and unreliable components is referred to as the missing data mask.

Calculating the noise estimate $\hat{N}(\tau, i)$ depends on the noise type. For babble noise, the method used in [4], where frames $\boldsymbol{Y}(\tau)$ classified as non-speech are temporally smoothed to produce the noise estimate $\hat{\boldsymbol{N}}(\tau)$, is applied. For impulsive noise, the impulses are first located in time and the noise estimates defined 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. Detecting the impulse locations is based on modelling speech as an autoregressive process and analysing the model prediction error as proposed in [6]. The parameters used in mask estimation are optimised using the development data described in Section 3.2.

2.2. GMM-based reconstruction

In missing-feature methods, the feature components $Y_r(\tau, i)$ classified as reliable are assumed fair estimates for the clean speech so that $X_r(\tau, i) \approx Y_r(\tau, i)$ whereas the unreliable components $Y_u(\tau, i)$ only provide an upper bound for the corresponding clean-speech values, $X_u(\tau, i) \leq Y_u(\tau, i)$. The features $X_u(\tau, i)$ are essentially missing and must be compensated for in speech recognition. In this work, the missing features are replaced with estimates that are calculated based on a clean speech model and reliable components within a fixed-length window. Other approaches for missing-feature reconstruction are discussed in Section 2.3.

2.2.1. Model training

The log-compressed mel-spectral data is processed in T-frame windows with each window represented as a TD-dimensional vector $\boldsymbol{x}(\tau)$, where D is the number of mel-spectral channels in the original data. The vectors $\boldsymbol{x}(\tau)$ are assumed independent and identically distributed (*i.i.d.*) and their distribution modelled as a Gaussian mixture model (GMM). Each mixture component m is associated with a full covariance matrix $\boldsymbol{\Sigma}(m)$ as proposed in [2]. The cluster indices m are modelled as a hidden variable, and the clusters and distribution parameters jointly

Copyright © 2011 ISCA

optimised using the expectation-maximisation (EM) algorithm.

While full-covariance models achieve high performance rates with a low number of Gaussian components, the number of model parameters increases quadratically with respect to T. In this work, maximum likelihood factor analysis is applied in model training to limit the number of parameters and possibly prevent overfitting. Factor analysis is based on modelling the covariance structure of high-dimensional data $\boldsymbol{x}$ using a lower-dimensional latent variable $\boldsymbol{z}$. The observations are modelled as

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{u}, \quad (1)$$

where $\boldsymbol{W}$ denotes the mapping from observation domain to factor domain, and $\boldsymbol{u}$ represents observation noise. The latent variable $\boldsymbol{z}$ is assumed Gaussian distributed with zero mean and unit variance, $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$, and the noise $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a diagonal covariance matrix. Equation (1) defines a Gaussian distribution for $\boldsymbol{x}$.

In order to achieve local dimensionality reduction within each GMM cluster, a mixture of factor analysers (MFA) [5] is used. The generative distribution for observations $\boldsymbol{x}$ in the MFA framework is given as

$$p(\boldsymbol{x}) = \sum_m c(m)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}(m), \boldsymbol{W}(m)\boldsymbol{W}(m)^{\mathsf{T}} + \boldsymbol{\Psi}), \quad (2)$$

where $c(m)$ denotes the weight of the mth mixture component. The noise covariance is assumed cluster-independent, $\boldsymbol{\Psi}(m) = \boldsymbol{\Psi}$. Equation (2) defines a GMM distribution with mean vectors $\boldsymbol{\mu}(m)$ and covariance matrices $\boldsymbol{\Sigma}(m) = \boldsymbol{W}(m)\boldsymbol{W}(m)^{\mathsf{T}} + \boldsymbol{\Psi}$. In this work, the clusters and model parameters $\boldsymbol{\mu}(m)$, $\boldsymbol{W}(m)$, and $\boldsymbol{\Psi}$ are jointly estimated using the EM algorithm proposed in [5].

2.2.2. Reconstruction

Given a TD-dimensional noisy observation vector $\boldsymbol{y}(\tau)$ divided in reliable and unreliable components $\boldsymbol{y}_r(\tau)$ and $\boldsymbol{y}_u(\tau)$ and a clean speech model with parameters Λ, the missing features $\boldsymbol{x}_u(\tau)$ can be estimated as a weighted sum of cluster-conditional maximum likelihood (ML) estimates as proposed in [2],

$$\hat{\boldsymbol{x}}_u(\tau) = \sum_m P(m|\boldsymbol{y}(\tau), \Lambda) E[\boldsymbol{x}_u|\, \boldsymbol{y}(\tau), \Lambda, m], \quad (3)$$

where $\boldsymbol{y}(\tau)$ denotes $\boldsymbol{x}_r = \boldsymbol{y}_r(\tau)$ and $\boldsymbol{x}_u \leq \boldsymbol{y}_u(\tau)$. The posterior probability for cluster m is calculated based on the prior probability $c(m)$ estimated from training data and the likelihood $p(\boldsymbol{x}_r = \boldsymbol{y}_r(\tau), \boldsymbol{x}_u \leq \boldsymbol{y}_u(\tau)|\Lambda, m)$. The likelihoods are calculated using diagonal covariances. The cluster-conditional estimates $E[\boldsymbol{x}_u|\, \boldsymbol{y}(\tau), \Lambda, m]$ cannot be solved in closed form, but computing the exact solution requires using iterative methods [2].

In this work, the cluster-conditional estimates are approximated as the minimum of the observed value and the unbounded estimate $E[\boldsymbol{x}_u|\boldsymbol{y}_r(\tau), \Lambda, m]$ that does not use the unreliable observations. Thus, the mth cluster-conditional estimate in Equation (3) is calculated as

$$E[\boldsymbol{x}_u|\boldsymbol{y}(\tau), \Lambda, m] = \min\{\boldsymbol{y}_u(\tau), E[\boldsymbol{x}_u|\boldsymbol{y}_r(\tau), \Lambda, m]\}, \quad (4)$$

where the $\min$ operator denotes a component-wise minimum and $E[\boldsymbol{x}_u|\boldsymbol{y}_r(\tau), \Lambda, m]$ is the mth unbounded estimate calculated as

$$E[\boldsymbol{x}_u|\boldsymbol{y}_r(\tau), \Lambda, m] = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{ur}\boldsymbol{\Sigma}_{rr}^{-1}(\boldsymbol{y}_r(\tau) - \boldsymbol{\mu}_r), \quad (5)$$

where $\boldsymbol{\mu}_r$ and $\boldsymbol{\mu}_u$ are the expected values of the reliable and unreliable components $\boldsymbol{x}_r$ and $\boldsymbol{x}_u$, $\boldsymbol{\Sigma}_{ur}$ their cross-covariance, and $\boldsymbol{\Sigma}_{rr}$ the covariance matrix of $\boldsymbol{x}_r$. The mean and covariance parameters in Equation (5) are subsets of the mean and covariance parameters of the mth Gaussian component in the clean speech model. Calculating the estimates iteratively as proposed in [2] was found too time consuming when the window width T increased.

2.2.3. Windowed data

In this work, the log-mel-spectral features are processed in windows spanning T frames. The frame shift between two consecutive windows is 1 frame in all the reported experiments. Since the windows overlap in time when the window width $T > 1$, each frame $\boldsymbol{X}(\tau)$ will be associated with several clean speech estimates. These are subvectors of $\hat{\boldsymbol{x}}(\tau)$ calculated as described in Section 2.2.2. In this work, estimates $\hat{\boldsymbol{X}}(\tau)$ are calculated as the average of the estimates for $\boldsymbol{X}(\tau)$ in the different windows.

2.3. Related methods

In this work, the GMM framework proposed for missing-feature reconstruction in [2] is evaluated on features that span several frames. Motivation for using windowed data comes from the recently introduced sparse imputation (SI) method [3]. This is a non-parametric method based on modelling the clean speech windows as linear combinations of a limited number of example windows stored in a clean speech dictionary. In experiments using large vocabulary continuous speech data, the optimal window width for SI in missing-feature reconstruction was found $T = 5 \ldots 20$ [4].

Other approaches that use time context in missing-feature reconstruction include correlation-based reconstruction [2] and hidden Markov model (HMM) based reconstruction [7]. The two methods are not, however, considered in this work as 1) using correlation-based reconstruction on realistic data has invariably resulted in lower performance scores than using GMM-based reconstruction and 2) using HMM is computationally heavy unless approximations such as proposed in [7] are applied. Moreover, using correlation across time resulted in improvements over frame-based HMM performance in some, but not all, noise conditions [7].

3. Experiments

3.1. 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 HMMs for context-dependent triphones constructed using phonetic decision trees. The output densities of the states are GMMs and the state durations are modelled as 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. Since all words and word forms can be represented with the statistical morphs, the decoding vocabulary is in practice unlimited [8].

Missing features are reconstructed in the 21-dimensional log-compressed mel-spectral domain. The mask thresholds γ

for missing-feature reconstruction in babble noise are 4 dB, 5 dB, and 6 dB for windows of 1, 5, and 10 frames, respectively, and $\gamma = 5$ dB for sparse imputation. The mask threshold for impulsive noise is $\gamma = 1$ dB. The clean speech models used in this work have 5 Gaussian components. The number was chosen as a reasonable compromise between performance and computation time. Gaussian mixtures were trained using the EM algorithm implemented in GMMBAYES Matlab toolbox and mixtures of factor analysers with a 50-dimensional factor domain estimated using a Matlab implementation[1] of the EM algorithm proposed in [5]. A fuzzy c-means algorithm from the GMMBAYES toolbox was used for initialisation in both cases. SI was used as described in [4]. The window width for SI ($T = 15$) and the mask thresholds γ for all models were optimised using the development data described in Section 3.2.

3.2. Data

The clean speech data used in this work is selected from the Finnish SPEECON database. Acoustic models are trained with a 30-hour training set that contains clean speech recorded with a headset in quiet conditions. The utterances include isolated words, sentences, and spontaneous speech. The exemplar dictionary for sparse imputation is sampled from the 14 hours of read sentences in the SPEECON training data and the GMM models are trained with 500 sentences (52 minutes) randomly selected from this subset. The speech–non-speech classifier used in mask estimation (Section 2.1) is trained with babble-noise corrupted television news data from the Finnish Broadcasting Company (YLE).

The missing-feature reconstruction methods are evaluated on 1118 clean speech utterances (113 minutes) 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 evaluation set includes utterances from 40 speakers. The development data contains 350 clean speech utterances (36 minutes) from 39 speakers corrupted with SNR 10 dB babble noise or SNR 0 dB impulsive noise i.e. hammering recorded 30 cm from the noise source (metal hammer on wood). All utterances are read sentences from the Finnish SPEECON database, and the training, evaluation, and development sets contain data from different speakers.

3.3. Results

Speech recognition results on babble-noise corrupted data are reported in Table 1 and results on impulsive-noise corrupted data in Table 2. We compare the performance of (a) the frame-based GMM, (b) GMM trained on 5-frame windows, (c) MFA trained on 5-frame windows, (d) GMM trained on 10-frame windows, and (e) MFA trained on 10-frame windows. Sparse imputation (SI) results are provided for reference. The speech recognition performance is reported in letter error rate (LER) which is better suited for Finnish than word error rate (WER). Statistical significance is evaluated in pairwise comparisons using Wilcoxon signed rank test at significance level $\alpha = 0.05$. For the purpose of statistical testing, each speaker-specific LER is considered an observation.

For babble noise (Table 1), the relative error reduction from using 5-frame windows (b) is 5–15 % compared to the frame-based GMM baseline (a). The improvements are statistically significant with $p < 0.0005$ in all pairwise comparisons. The relative error reduction from using 10-frame windows (d) is less than 5 % compared to the frame-based GMM (a) at all noise levels, and the differences in LER are not statistically significant at some SNRs. Using MFA consistently improves the recognition performance compared to GMM, but the differences between systems (b) and (c) or (d) and (e) are not statistically significant at some SNRs. The overall relative error reduction from using MFA trained on 5-frame windows (c) is 6–22 % compared to the frame-based GMM (a). The error reduction is 13–22 % at SNR 5–15 dB.

[1] Available in learning.eng.cam.ac.uk/zoubin/software.html

Table 1: Letter error rates for babble noise.

	SNR 15	SNR 10	SNR 5	SNR 0
(a) GMM–T1	8.5	23.2	56.2	78.8
(b) GMM–T5	7.7	19.5	48.9	74.5
(c) MFA–T5	7.4	18.1	47.0	74.3
(d) GMM–T10	8.4	22.2	54.9	79.4
(e) MFA–T10	7.9	21.0	53.5	79.0
SI	7.1	16.4	43.4	72.3

Table 2: Letter error rates for impulsive noise.

	SNR 5	SNR 0	SNR -5	SNR -15
(a) GMM–T1	26.1	30.7	36.8	48.9
(b) GMM–T5	19.3	25.7	33.1	47.1
(c) MFA–T5	18.4	24.5	32.3	45.7
(d) GMM–T10	12.8	19.5	28.8	45.4
(e) MFA–T10	12.5	18.0	27.0	43.6
SI	18.7	24.7	32.9	46.2

For impulsive noise (Table 2), the relative error reduction from using 5-frame windows (b) is 4–26 % and from using 10-frame windows (d) 7–51 % compared to the frame-based GMM baseline (a). The improvements are statistically significant with $p < 0.0001$ in all pairwise comparisons. Using MFA consistently improves the recognition performance compared to GMM, although the difference between systems (d) and (e) is not statistically significant at SNR 5 dB. The improvements observed in other conditions are statistically significant with $p < 0.001$ in all pairwise comparisons, and the best results are achieved with MFA trained on 10-frame windows (e). The relative error reduction from using system (e) is 10–52 % compared to the frame-based GMM (a).

4. Discussion

4.1. Main results

We have investigated using GMM-based missing-feature reconstruction on multi-frame windows and applied MFA to limit the number of parameters in the clean speech model. Results on the noisy speech recognition task indicate that both window-based processing and using MFA generally improve the performance of the method, although the difference between GMM and MFA performance was not statistically significant in some noise conditions. Adding time context resulted in relative error reduc-

tions up to 22 % in babble noise condition and up to 52 % in impulsive noise condition. Using MFA resulted in relative error reductions of 2–8 % depending on the noise type and SNR. In the babble noise condition, best results were achieved using SI, and in the impulsive noise condition, using MFA trained on $T = 10$ frame windows.

4.2. Noise types

In this work, methods were evaluated on clean speech data corrupted with babble and impulsive noise, with different results depending on the noise type. In the babble-noise condition, error rates obtained with the best GMM-based and SI approaches were quite similar, whereas in the impulsive-noise condition, there was a notable difference between the GMM-based systems trained on $T = 10$ frame windows and SI. This is possibly because in the impulsive-noise condition, the masked clean speech values may include time–frequency components with relatively high energies, which is unlikely in the babble-noise condition. The location of high-energy components in a speech segment or window is likely among the most reliable cues for differentiating between different speech tokens, which is relevant when using a method like SI that directly matches the reliable observations with clean speech examples. Consequently, such methods may be more sensitive to missing the high-energy components than statistical methods like GMM-based reconstruction.

Noise type was also observed to affect the optimal window width for GMM-based reconstruction. The best results were achieved with 5-frame windows when applied on babble noise, whereas in impulsive noise, using 10-frame windows further improved the GMM-based reconstruction performance. Increasing the window width is not without disadvantage as models trained on higher-dimensional data need to represent a wider variation of features within each cluster. Such models may be prone to misclassification especially when windows contain few reliable values. Since impulsive noise corruption is localised in time, increasing the window width generally increases the percentage of reliable values within a window, whereas in babble noise, the percentage of reliable values increases very little on average. Finally, we note that using $T = 10$ frame windows in GMM-based reconstruction increased the number of deletion errors in the babble noise condition. Similar result has been reported for SI when the window width is increased past optimum range [4].

4.3. Future work

The GMM-based approach used in this work did not attempt to calculate bounded ML estimates as proposed in [2], but unbounded estimates were simply restricted not to exceed the observed values. While the solution is computationally efficient, better results may be achieved with proper bounding. The increase in computation time should remain acceptable with, for example, the non-iterative method proposed for computing approximate bounded MMSE estimates in [9]. Another imperfection in the reported experiments concerns comparing the GMM and MFA approaches. Parametric models are usually compared in a setting with a fixed number of model parameters, which would allow more clusters for MFA. The dimension of the factor domain should also be optimised for improved performance.

5. Acknowledgements

This work was supported in part by the Helsinki Graduate School in Computer Science and Engineering and the Academy of Finland in projects No 129674 and 140299. The authors would like to thank Jort F. Gemmeke for the sparse imputation code.

6. References

[1] 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, 2001.

[2] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," Speech Communication, vol. 43, no. 4, pp. 275–296, 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. STSP, vol. 4, no. 2, pp. 272–287, 2010.

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

[5] Z. Grahmani and G. E. Hinton, "The EM Algorithm for mixtures of factor analyzers", Technical Report CRG-TR-96-1, University of Toronto, 1996.

[6] S. V. Vaseghi and P. J. W. Rayner, "Detection and suppression of impulsive noise in speech communication systems", IEE Proceedings, vol. 137, no. 1, pp. 38–46, 1990.

[7] B. J. Borgström and A. Alwan, "HMM-based reconstruction of unreliable spectrographic data for noise robust speech recognition", IEEE Trans. Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1612–1623, 2010.

[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", Computer Speech and Language, vol. 20, no. 4, pp. 515–541, 2006.

[9] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models: a reconstruction approach to partly occluded features," in Proc. ICASSP, Taipei, Taiwan, 2009, pp. 3869–3872.