

A Hidden Markov Model-Based Missing Data Imputation Approach

Yu Luo, Limin Du

Center for speech interactive information research
Institute of Acoustics, Chinese Academy of Sciences
luoy@iis.ac.cn, dulm@iis.ac.cn

Abstract

The accuracy of automatic speech recognizer degrades rapidly when speech was distorted by noise. Robustness against noise arises to be one of the challenge problems. In this paper, a hidden Markov model (HMM) based data imputation approach is presented to improve speech recognition robustness against noise at the front-end of recognizer. Considering the correlation between different filter-banks, the approach realizes missing data imputation by a HMM of L states, each of which has a Gaussian output distribution with full covariance matrix. "Missing" data in speech filter-bank vector sequences are recovered by MAP procedure from local optimal state path or marginal Viterbi decoded HMM state sequence.

The potential of the approach was tested using speaker independent continuous mandarin speech recognizer with syllable-loop of perplexity 402 for both Gaussian and babble noises each at 6 different SNR levels ranging from 0dB to 25dB, showing a significant improvement in robustness against additive noises.

1. Introduction

The performance of modern Automatic Speech Recognition (ASR) systems degrades rapidly when clean speech was distorted by additive noise. To improve ASR system's robustness against additive noise, Missing data methods [8][9][10] are developed in literature. Missing data methods assume that additive noise distort speech differently in different time-spectrum region. The noise greatly distorted speech regions, which has lower local SNR, are marked as "missing" and the noise slightly distorted speech regions, which has higher local SNR, are marked as "reliable". After mask estimation, speech recognition can be done either using "reliable" data (marginal model method) or using recovered data (data imputation method). Missing data method doesn't make an assumption about the characteristic of additive noise and results in a potential of improving the robustness of ASR system against non-stationary noise.

Several data imputation method are studied in literature, such as Single Gauss Model set based Data Imputation [1], Gaussian Mixture Model based data imputation [2] or Vector Quantization based data imputation [3]. Most of these methods process each frame independently and fails to take account of the time evolution of the spectrum parameters.

The time evolution of spectrum parameters pays an important role in speech recognition. Continuous density HMM which state distributions are M multivariate-Gaussians with diagonal covariance matrices is introduced in literature [4]. Based on the assumption that "missing" components are independent of "reliable" components, HMM state sequences

are estimated and missing components are reconstructed by state based data imputation method. However, this assumption is doubtful, because there are overlaps between neighbor filters in filter-bank analysis. So there is obvious correlation between neighbor filter-banks. To recover "missing" components more accurately, we must consider the correlation between different filter-banks.

In this paper, we use HMM to model the time evolution of filter-bank vector sequence and full covariance matrix is introduced to represent the correlation between different filter-banks. Each state distribution of HMM is selected as single Gaussian with full covariance matrix. Assuming that speech feature vector sequences come from such an L state full covariance matrices HMM, Local Optimal Path procedure and marginal Viterbi decoding process are used to estimate the HMM state sequence. Then, according to state distribution, "missing" data is recovered by MAP procedure.

To evaluate the potential of HMM-based Data Imputation method, the robust performance of HMM-based Data Imputation method was tested at different SNR level (ranged from 0dB to 25 dB) in a complex task, speaker independent continuous mandarin speech recognition with syllable-loop of perplexity 402.

We discuss HMM-based Data Imputation method in section 2. Section 3 gives ideal mask estimation, which will be used to evaluate HMM-based data imputation method. Experimental result is presented and discussed in section 4 and conclusion is given in section 5.

2. Hidden Markov Model-based Data Imputation

After mask estimation, each speech feature vector S is split into two vectors, S^m and S^o . S^m represents "missing" components of S and S^o represents "reliable" components of S . We try to estimate the "missing" vector sequence $[S_1^m, S_2^m, \dots, S_T^m]$ by the "reliable" vector sequence $[S_1^o, S_2^o, \dots, S_T^o]$ and the HMM parameter $\lambda = [a, A, B]$.

Hidden Markov Model-based data imputation is carried out in two steps:

First, according to $[S_1^o, S_2^o, \dots, S_T^o]$ and $\lambda = [a, A, B]$, estimate HMM state sequence $X=[x_1, x_2, \dots, x_T]$;

Second, at each time t , estimate the "missing" vector S_t^m according to x_t and S_t^o .

2.1. Hidden Markov Model (HMM) [5]

Considering a system, which may be at one of a set of L distinct states at any time t , we use symbol $[Q_1, \dots, Q_L]$ to represent these states. At a discrete time t , the state of the system is represented as x_t ($t=1, 2, \dots, T$)

$$x_t \in [Q_i] \quad (1 \leq i \leq L, 1 \leq t \leq T) \quad (1)$$

At a discrete time t , system state x_t is determined by initial state distribution $a=[a_1, \dots, a_L]$ and state-transition matrix $A=[A_{ij}]$

$$a_i = P(x_1 = Q_i) \quad (1 \leq i \leq L) \quad (2)$$

$$A_{ij} = P(x_t = Q_j / x_{t-1} = Q_i) \quad (t > 1, 1 \leq i, j \leq L) \quad (3)$$

$b_i(S)$, the state distribution of observation vector at state i , is represented as

$$b_i(S) = P_{Q_i}(S) = P(S/x = Q_i) \quad (1 \leq i \leq L) \quad (4)$$

$B=[b_1(S), \dots, b_L(S)]$. A HMM can be represented by parameter $\lambda = [a, A, B]$.

2.2. Local Optimal state Path procedure-based Data Imputation

Local Optimal state Path procedure can be used to estimate system state sequence. Then, “missing” components will be recovered. We call this method Local Optimal state Path-based Data Imputation (LOPDI). LOPDI is carried out as following steps:

1) Initialization

At time 1, system state x_1 is initialized according to initial state distribution, a , “reliable” vector, S_1^o , and the conditional probability, $b_j(S_1^o)$

$$x_1 = \arg \max_{Q_j} [a_j b_j(S_1^o)] \quad (5)$$

where $b_j(S_1^o)$ represents the conditional probability of observing “reliable” vector S_1^o on condition that system is at state Q_j . $b_j(S_1^o)$, the conditional probability, is calculated by marginal process

$$b_j(S_1^o) = P_{Q_j}(S_1^o) = \int P_{Q_j}(S_1^o S^m) dS^m \quad (6)$$

2) Local Optimal state path estimation

For each time $t > 1$, current state x_t is determined by the last state $x_{t-1} = Q_i$, transition probability a_{ij} and current “reliable” vector S_t^o

$$x_t = \arg \max_{Q_j} [a_{ij} * b_j(S_t^o)] \quad 2 \leq t \leq T \quad (7)$$

where $b_j(S_t^o)$ represents the conditional probability of observing “reliable” vector S_t^o on condition that system is at state Q_j . $b_j(S_t^o)$, the conditional probability, is calculated by marginal process

$$b_j(S_t^o) = P_{Q_j}(S_t^o) = \int P_{Q_j}(S_t^o S^m) dS^m \quad (8)$$

3) Recover “missing” components

At time t , estimate the “missing” vector S_t^m so that $b_{x_t}(S_t^o S_t^m)$, the probability of observing speech feature vector $S=[S_t^o S_t^m]$ at state x_t , is maximized

$$\hat{S}_t^m = \arg \max_{S^m} (b_{x_t}(S_t^o S_t^m)) \quad (9)$$

If the state distribution of observation vector is Gaussian

$$b_i(S) = P_{Q_i}(S) = \frac{\exp \left\{ -\frac{1}{2} (S - \mu_i)^T \theta_i^{-1} (S - \mu_i) \right\}}{(2\pi)^{\frac{n}{2}} |\theta_i|^{-\frac{1}{2}}} \quad (10)$$

According to equation (9), “Missing” components are recovered by [1]

$$\hat{S}_t^m = \mu_{x_t, m} + \theta_{x_t, mo} \theta_{x_t, oo}^{-1} (S^o - \mu_{x_t, o}) \quad (11)$$

where the $\mu_{x_t, m}$ represent the mean vector of “missing” components, $\mu_{x_t, o}$ is the mean vector of “reliable” components, $\theta_{x_t, oo}$ is the auto-covariance matrix of “reliable” components, $\theta_{x_t, mo}$ is the cross-covariance matrix of “missing” and “reliable” filter-bank.

Local Optimal state Path procedure has the advantage that we can estimate current HMM state and recover the speech vector in real time. But it may fall into local optimal state and output a sequence of incorrect states. Estimation error increases rapidly in this situation.

2.3. Marginal Viterbi decoding process-based Data Imputation

To estimate the optimal state sequence, we introduce marginal Viterbi decoding process [5]. We call this method marginal Viterbi decoding process-based data imputation (VITDI).

To do marginal Viterbi decoding process, define

$$\varphi_t(i) = P[x_1, \dots, x_{t-1}, x_t = Q_i, S_1^o, \dots, S_t^o | \lambda] \quad (12)$$

where $\varphi_t(i)$ represents the highest probability of observing $[S_1^o, S_2^o, \dots, S_t^o]$ along all possible state paths and system is at state i at time t . if $\varphi_t(i)$ is given, $\varphi_{t+1}(j)$ can be estimated by

$$\begin{aligned} \varphi_{t+1}(j) &= P[x_1, \dots, x_t, x_{t+1} = Q_j, S_1^o, \dots, S_t^o, S_{t+1}^o | \lambda] \\ &= \left\{ \max_i [\varphi_t(i) A_{ij}] \right\} \cdot P_{x_{t+1}=Q_j}(S_{t+1}^o) \\ &= \left\{ \max_i [\varphi_t(i) A_{ij}] \right\} \cdot b_j(S_{t+1}^o) \end{aligned} \quad (13)$$

To keep track of the optimal state sequence, define

$$\psi_{t+1}(j) = \arg \max_i [\varphi_t(i) A_{ij}] \quad (14)$$

VITDI is carried out as following steps:

1) Initialization

$$\begin{aligned} \delta_1(i) &= a_i b_i(S_1^o) \quad (1 \leq i \leq L) \\ \psi_1(i) &= 0 \end{aligned} \quad (15)$$

where a_i is initial state probability. $b_i(S_1^o)$ represents the conditional probability of observing “reliable” vector S_1^o on condition that system is at state Q_i at time 1. $b_i(S_1^o)$ is calculated by marginal process.

2) Recursion

$$\begin{aligned} \varphi_t(j) &= \left\{ \max_i [\varphi_{t-1}(i) A_{ij}] \right\} \cdot b_j(S_t^o) \quad (2 \leq t \leq T, 1 \leq j \leq L) \\ \psi_t(j) &= \arg \max_i [\varphi_{t-1}(i) A_{ij}] \quad (2 \leq t \leq T, 1 \leq j \leq L) \end{aligned} \quad (16)$$

where A_{ij} is transition probability (from state i to state j). $b_j(S_t^o)$ represents the conditional probability of observing “reliable” vector S_t^o on condition that system is at state Q_j at time t . $b_j(S_t^o)$ is calculated by marginal process.

3) Terminate recursion at time T

$$\begin{aligned} P^* &= \max_{1 \leq i \leq L} \varphi_T(i) \\ q_T^* &= \arg \max_{1 \leq i \leq L} \varphi_T(i) \end{aligned} \quad (17)$$

4) State path backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \quad (t = T-1, \dots, 1) \quad (18)$$

5) Recover “missing” components

$$\hat{S}_t^m = \mu_{x,m} + \theta_{x,m} \theta_{x,o}^{-1} (S_t^o - \mu_{x,o}) \quad (19)$$

According to $[S_1^o, S_2^o, \dots, S_T^o]$, LOPDI estimates the optimal state sequence $[x_1, x_2, \dots, x_T]$ by marginal Viterbi decoding process. So VITDI is expected to recover “missing” components better than LOPDI method.

3. Mask Estimation

In this paper, Mask estimation and data imputation will be done in 26-dimension Mel-filter-bank vector domain. 26 triangular filters equally spaced along the Mel-scale are used to do filter-bank analysis.

We create noisy speech by adding noise to “clean” speech (global SNR=0, 5, 10, 15, 20, 25dB). Clean speech and additive noise are saved for ideal mask estimation. If pre-reserved clean speech vector is given as S and additive noise vector is given as N. Ideal mask estimation is done by

$$MSK_i(k) = \begin{cases} 1 & \text{if } S\hat{N}R_i(k)=10 \text{Log}_{10} \left(\frac{S_i(k)}{N_i(k)} \right) > \delta \\ 0 & \text{if } S\hat{N}R_i(k)=10 \text{Log}_{10} \left(\frac{S_i(k)}{N_i(k)} \right) \leq \delta \end{cases} \quad (20)$$

where $S_i(k)$ represents the energy of the k-th sub-band of the i-th Mel-filter-bank of clean speech. $N_i(k)$ represents the energy of the k-th sub-band of the i-th Mel-filter-bank of additive noise. $MSK_i(k)=1$ means that the k-th sub-band is “reliable” and $MSK_i(k)=0$ means that the k-th sub-band is “missing”. δ , the threshold of local SNR, is determined by auditory masking effect. In this paper, we select $\delta = -5\text{dB}$.

Ideal Mask Estimation is hardly used for a practical application, for it is a very difficult problem to accurately separate speech signal from noise. In this paper, we use ideal mask estimation to evaluate the potential of Hidden Markov Model-based data imputation algorithms.

4. Experimental result and discussion

We carried out ASR Experiment to compare LOPDI and VITDI with Single Gauss Model set based Data Imputation (SGMDI).

5.1 Experiment condition

We use HTK3.0 [6] build a speaker independent continuous mandarin speech recognition system.

The training and testing data comes from 863-speech corpus, which is open and widely used in mandarin speech recognition. Speech data of 158 persons (79 male, 79 female) is used for training and data of left 8 persons (4 male, 4 female) is used for testing.

Two typical additive noises (gauss white noise, babble noise, from NoiseX-92 database) are used in ASR experiments. We create noise-distorted speech by adding noise to “clean” speech (global SNR=0, 5, 10, 15, 20, 25dB). “Clean” speech and additive noise are saved for ideal mask estimation.

To model the continuous speech, triphone model is used in HMM based ASR system. Each triphone model has 3 states, and each state uses 7 diagonal-covariance mixture gauss. There are over 8000 states in our speech recognition system.

Speech feature vectors are 39-dimension MFCC_E_D_A vectors. The waveform speech is transformed to Mel-filter-bank first. Mel-filter-bank features were computed every 10ms

using a 25ms hamming window. 26 Mel-spaced triangular filters, ranging from 0Hz to 8000Hz, are used to calculate the Mel-filter-bank features. Mask estimation and data imputation will be done in Mel-filter-bank vector domain. After data imputation, Mel-filter-bank vectors are transformed to MFCC_E_D_A and sent to HMM-based recognizer.

HMM-based recognizer does Viterbi decoding with mandarin syntax independent all-syllable-loop of perplexity 402.

The HMM which is used to do data imputation by LOPDI and VITDI algorithm contains 256 states ($L=256$), and the Single Gauss Model set used by SGMDI algorithm contains 256 Gaussians ($N=256$).

5.2 Experimental results

Experimental results of large vocabulary speaker independent continuous Chinese speech distorted by 2 typical additive noises (gauss white noise, babble noise) are presented and discussed in this section.

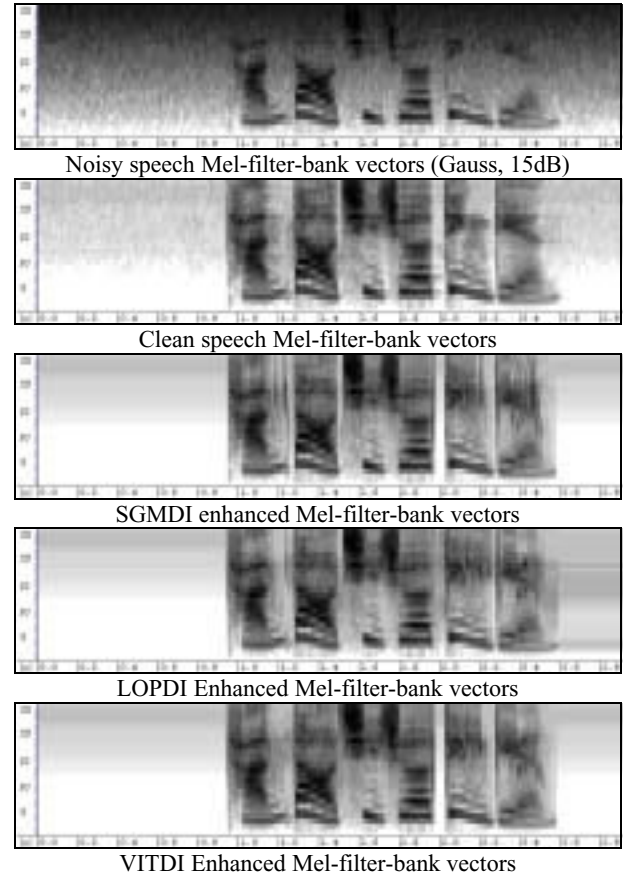


Figure 1: Enhanced Mel-filter-banks for gauss white noise distorted mandarin speech (“tan2 dao4 qi4 che1 ding4 dian3”), SNR=15dB

Additive noise distorted distribution of Mel-filter-bank vectors of clean speech. After ideal mask estimation and data imputation, recovered Mel-filter-bank vectors are more similar to Mel-filter-bank vectors of clean speech than those of noise-distorted speech.

SGMDI process each frame independently and fails to take account of the time evolution of the spectrum parameters. We

use HMM to model the distribution and dynamics of speech feature vectors. HMM-based Data Imputation can reduce the discontinuities between neighbor Mel-filter-banks, as shown in figure 1.

LOPDI is unable to get the optimal state sequence. It may fall into local optimal but incorrect states and output a sequence of mis-estimated states. To overcome this shortcoming, VITDI uses marginal Viterbi decoding process to estimate the optimal state sequence. So it recovers “missing” components better than LOPDI, as shown in figure 1.

After mask estimation and data imputation, Mel-filter-banks are transformed to MFCC_E_D_A, and sent to HMM-based recognizer. Word (mandarin syllable) accuracy [6] of ASR experiments is given in table 1.

Table 1: ASR experimental results

Noise type	SNR (dB)	Noisy (%Acc.)	SGMDI (%Acc.)	LOPDI (%Acc.)	VITDI (%Acc.)	Clean (%Acc.)
Gauss Noise	0	2.49%	3.56%	-13.43%	6.30%	74.71%
	5	2.34%	20.24%	12.96%	26.58%	
	10	1.31%	39.49%	35.72%	42.78%	
	15	12.09%	51.02%	55.89%	55.90%	
	20	28.00%	60.17%	61.57%	63.20%	
	25	56.14%	67.25%	68.53%	69.26%	
Babble Noise	0	-3.07%	13.64%	-8.34%	18.08%	
	5	-5.81%	32.03%	14.51%	36.58%	
	10	8.15%	51.23%	38.76%	54.46%	
	15	29.74%	62.36%	62.63%	64.37%	
	20	45.97%	68.75%	67.05%	70.07%	
	25	58.91%	71.49%	72.37%	72.64%	

ASR experiments show that the performance of a speech recognizer with a complex acoustic model is badly degraded by additive noise. It also shows that different noise has different affect upon clean speech. In our experiments, ASR system performs better in babble noise environment when SNR>5dB.

Data imputation methods can dramatically improve ASR system’s robustness against additive noise, as shown in table 1. LOPDI can improve system performance in gauss noise environment with high SNR (SNR>=15dB). When SNR<=10dB, the performance of LOPDI degrade rapidly. This can be explained by the fact that more and more filter-banks are marked as “missing” with decrease of global SNR. When most components of neighbor filter-bank vectors are marked as “missing”, it’s unreliable to estimate the state sequence by local optimal path procedure.

To estimate the optimal state sequence, marginal Viterbi decoding process is introduced by VITDI. In both noisy conditions (Gaussian noise, babble noise), VITDI performs better than LOPDI and SGMDI at all SNR level, as shown in table 1.

5. Conclusion

In this paper, a hidden Markov model (HMM) based data imputation approach is presented to improve speech recognition robustness against noise at the front-end of recognizer. We use local optimal path procedure and Viterbi decoding process to estimate the HMM state sequence. Then, According to the distributions of each state, “missing” data is recovered by MAP procedure. LOPDI can improve ASR system’s robustness against additive noise in high SNR

environment, but fails in low SNR environment. Marginal Viterbi decoding process is introduced in VITDI to estimate the optimal state sequence. Experimental result shows that VITDI performs better than LOPDI and SGMDI at all SNR level.

6. Acknowledgement

This research is supported by 973 national key fundamental research project, “Chinese spontaneous speech dialog theory and experimental platform research”.

7. References

- [1] B. Raj, M. Seltzer, and R. Stern. Reconstruction of damaged spectrographic features for robust speech recognition. In Proc. ICSLP’00, volume 1, pages 375-360, Beijing, China, October 2000
- [2] Ph. Renevey and A. Drygajlo, "Statistical estimation of unreliable features for robust speech recognition", in Proceedings of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2000), Istanbul, Turkey, June 5-9, 2000, pp. 1731-1734.
- [3] Philippe Renevey, Rolf Vetter and Jens Kraus “Robust speech recognition using missing feature theory and vector quantization” Eurospeech 2001, Scandinavia, pp1107.
- [4] Ljubomir Josifovski, Martin Cooke, Phil Green, and Ascension Vizinho, “State based imputation of missing data for robust speech recognition and speech enhancement”. in Eurospeech vol. 6, pp. 2833–2836, 1999.
- [5] Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of speech Recognition”, Prentice hall PTR
- [6] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland, The HTK Book (for HTK Version 3.0)
- [7] A. Vizinho, P. Green, M. Cooke and L. Josifovski, Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study, Eurospeech’99, Budapest, 1999
- [8] Martin Cooke, Phil Green, Ljubomir Josifovski and Ascension Vizinho, “Robust ASR with unreliable data and minimal assumptions”. Robust 99, Tampere, Finland.
- [9] Morris,A.C., Cooke,M. & Green,P. (1998) "Some solutions to the missing feature problem in data classification, with application to noise robust ASR", Proc. ICASSP’98, pages 737-740.
- [10] Jon Barker, Ljubomir Josifovski, Martin Cooke and Phil Green “Soft decisions in missing data techniques for robust automatic speech recognition” ICSLP-2000, Beijing, pp373-376
- [11] Ph. Renevey and A. Drygajlo, "Introduction of a Reliability Measure in Missing Data Approach for Robust Speech Recognition", in Proceedings of 10th European Signal Processing Conference (EUSIPCO 2000), Tampere, Finland, Sept. 5-8, 2000, pp. 473-476.