Speech Recognition by Dereverberation Method Based on Multi-channel LMS Algorithm in Noisy Reverberant Environment

Kyohei Odani [*] (odani@spa.sys.eng.shizuoka.ac.jp)
Graduate School of Engineering, Shizuoka University, Japan
(Advisers: Longbiao Wang [†] and Atsuhiko Kai [‡])

1 Introduction

In a distant-talking environment, channel distortion drastically degrades speech recognition performance because of mismatches between the training and test environments. The current approaches focusing on robustness issues for automatic speech recognition (ASR) in noisy reverberant environments can be classified as speech enhancement, robust feature extraction, or model adaptation methods.

In this paper, we focus on speech enhancement in a distant-talking environment. Previously, Wang et al. [1] proposed a robust distant-talking speech recognition method based on power spectral subtraction (SS) employing the adaptive multi-channel least mean squares (MCLMS) algorithm. In their study, late reverberation was treated as additive noise, and a noise reduction technique based on power SS was proposed to estimate the power spectrum of clean speech using an estimated power spectrum of the impulse response. To estimate the power spectra of the impulse responses, they extended the MCLMS algorithm for identifying impulse responses in a time domain [2] to a frequency domain. We proposed a blind dereverberation method based on generalized SS (GSS), which has been shown to be effective for noise reduction, instead of power SS [3]. The dereverberation method based on GSS with beamforming achieved a relative word error reduction rate of 9.8% and 31.4% compared to the dereverberation method based on power SS with beamforming and the conventional cepstral mean normalization (CMN) with beamforming, respectively. However, both the power SS-based method [1] and GSS-based method [3] were evaluated in a simulated reverberant environment without additive noise.

In this paper, we evaluate the blind dereverberation methods in a real noisy reverberant environment with stationary noise or non-stationary noise. To suppress the stationary noise and non-stationary noise, we use a noise reduction technique based on GSS and a blind source separation based on independent component analysis (ICA), respectively. FastICA is one of the most popular algorithm for ICA. Our experimental result shows that combining the Efficient FastICA (EFICA) [4], which is an improved version of FastICA, is effective for our dereverberation method [6].

2 Outline of Dereverberation

The schematic diagram of our dereverberation method is shown in Fig. 1. The late reverberation are reduced from the spectrum of multi-channel distorted speech by our dereverberation method using the estimated spectrum of impulse response. Thereafter, the early reverberation is normalized by CMN at the feature extraction stage.

2.1 Dereverberation based on GSS

If speech $s[t]$ is corrupted by convolutional noise $h[t]$, the observed speech $x[t]$ becomes $x[t] = h[t] * s[t]$, where * denotes the convolution operation. If the length of the impulse response is much smaller than the analysis window length T used in the short-time Fourier transform (STFT), the STFT of the distorted speech equals that of the clean speech multiplied by the STFT of the impulse response $h[t]$. However, if the length of the impulse response is much greater than the analysis window size, the STFT of the distorted speech is usually approximated by

$$X(\tau,\omega) \approx S(\tau,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(\tau-d,\omega)H(d,\omega), \tag{1}$$

where τ is the frame index, $H(\omega)$ is the STFT of the impulse response, $S(\tau,\omega)$ is the STFT of the clean speech s, D is the number of reverberation windows, and $H(d,\omega)$ denotes the part of $H(\omega)$ corresponding to the frame delay d.

In [3], we proposed a dereverberation method based on GSS to estimate the STFT of the clean speech $\hat{S}(\tau,\omega)$ based on Eq. (1). To estimate the spectrum of the impulse response for the GSS, they extended the MCLMS algorithm for identifying the impulse responses in a time domain to a frequency domain. The estimated spectrum of clean speech may not be very accurate due to the estimation error of the impulse response, especially the estimation error of early part of the impulse response. In addition, the unreliable estimated spectra in previous frames cause a furthermore

[*] He is studying in the 2nd year at the course of Systems Engineering of this graduate school (Master course).
[†] He has been an associate professor at Nagaoka University of Technology.
[‡] He has been an associate professor in the Faculty of Engineering at Shizuoka University.

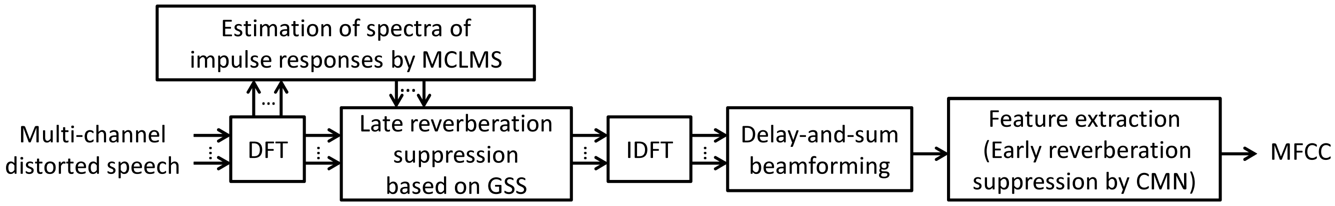

Estimation of spectra of impulse responses by MCLMS
Multi-channel distorted speech
DFT
Late reverberation suppression based on GSS
IDFT
Delay-and-sum beamforming
Feature extraction (Early reverberation suppression by CMN)
MFCC

Fig 1: Schematic diagram of our dereverberation method

estimation error in the current frame. In this paper, the late reverberation is reduced based on the GSS, while the early reverberation is normalized by CMN at the feature extraction stage.

Assuming that the phases of different frames are non-correlated for the sake of simplicity, the estimated spectrum $\hat{X}(\tau,\omega)$ obtained by reducing the late reverberation then becomes

$$|\hat{X}(\tau,\omega)|^{2n} \approx max\left\{|X(\tau,\omega)|^{2n} - \alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(\tau-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta \cdot |X(\tau,\omega)|^{2n}\right\}, \tag{2}$$

where $|\hat{X}(\tau,\omega)|^{2n} = |\hat{S}(\tau,\omega)|^{2n}|\hat{H}(0,\omega)|^{2n}$, $\hat{S}(\tau,\omega)$ is the spectrum of estimated clean speech, $\hat{H}(\tau,\omega)$ is the STFT of the impulse response obtained by frequency domain MCLMS algorithm in sec. 2.2, α is the noise over estimation factor, β is the spectral floor parameter to avoid negative or under flow values, and n is the exponent parameter, respectively.

2.2 Blind Estimation of Impulse Responses

In this section, we explain about blind estimation of the spectra of impulse response $\hat{H}(d,\omega)$ using Eq. (2). In [2], MCLMS in a time domain was proposed to blindly estimate the impulse responses of each channel. In this paper, we used a variable step-size unconstrained MCLMS (VSS-UMCLMS) algorithm extended in a time-domain to a frequency domain.

In the absence of additive noise, we have the following relation of correlation matrix and impulse response.

$$\mathbf{R}_{X_iX_i}(\tau+1)\mathbf{H}_j(\tau) = \mathbf{R}_{X_iX_j}(\tau+1)\mathbf{H}_i(\tau) i,j = 1,2,\cdots,N, i \neq j, \tag{3}$$

where

$$\begin{aligned}\mathbf{R}_{X_iX_j}(\tau) &= E[\mathbf{X}_i(\tau)\mathbf{X}_j^T(\tau)], &(4)\\ \mathbf{X}_i(\tau) &= [X_i(\tau), X_i(\tau-1),\cdots,X_i(\tau-D+1)]^T, &(5)\\ \mathbf{H}_i(\tau) &= [H_i(\tau,0),\cdots,H_i(\tau,d),\cdots,H_i(\tau,D-1)]^T, &(6)\end{aligned}$$

i is the channel number, $\mathbf{X}_i(\tau)$ is the spectrum of observed signal at frame τ, $\mathbf{H}_i(\tau)$ is the spectrum of impulse response at frame τ, $H_i(\tau,d)$ is the spectrum of impulse response at frame τ corresponding to the frame delay d.

Over all channels, we have written as Eq. (7) which transposed the right side and got together of Eq. (5).

$$\mathbf{R}_{X+}(\tau+1)\mathbf{H}(\tau) = \mathbf{0}, \tag{7}$$

where

$$\begin{aligned}\mathbf{R}_{X+}(\tau) &= \begin{bmatrix} \sum_{i\neq1}\mathbf{R}_{X_iX_i}(\tau) & -\mathbf{R}_{X_2X_1}(\tau) & \cdots & -\mathbf{R}_{X_NX_1}(\tau) \\ -\mathbf{R}_{X_1X_2}(\tau) & \sum_{i\neq2}\mathbf{R}_{X_iX_i}(\tau) & \cdots & -\mathbf{R}_{X_NX_2}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{X_1X_N}(\tau) & -\mathbf{R}_{X_2X_N}(\tau) & \cdots & \sum_{i\neq N}\mathbf{R}_{X_iX_i}(\tau) \end{bmatrix}, &(8)\\ \mathbf{H}(\tau) &= [\mathbf{H}_1(\tau)^T, \mathbf{H}_2(\tau)^T, \cdots, \mathbf{H}_N(\tau)^T]^T. &(9)\end{aligned}$$

Eq. (7) consists of using true impulse response in the absence of additive noise. When additive noise is presented or the estimated impulse response is used, however, the estimation error is observed like Eq. (10).

$$\tilde{\mathbf{R}}_{X+}(\tau+1)\hat{\mathbf{H}}(\tau) = \mathbf{E}(\tau+1), \tag{10}$$

where $\tilde{\mathbf{R}}_{X+}$ is the matrix of Eq .(9) calculated using noisy observed signal, $\mathbf{E}$ is the estimation error. $\hat{\mathbf{H}}$ is adaptively trained by minimizing the cost function obtained from the estimation error. The learning equation by unconstrained MCLMS is following as

$$\hat{\mathbf{H}}(\tau+1) = \hat{\mathbf{H}}(\tau) - 2\mu\tilde{\mathbf{R}}_{X+}(\tau+1)\hat{\mathbf{H}}(\tau), \tag{11}$$

where μ is the step-size. Multi-channel impulse responses can estimate by updating Eq. (11).

VSS-UMCLMS is the algorithm to determine the step-size μ of Eq. (11). The step-size μ is automatically updated like Eq. (12).

$$\mu_{opt}(\tau+1) = \frac{\hat{\mathbf{H}}^T(\tau)\Delta\mathbf{J}(\tau+1)}{||\Delta\mathbf{J}(\tau+1)||^2}, \tag{12}$$

where

$$\Delta\mathbf{J}(\tau+1) \quad \approx \quad \frac{2\tilde{\mathbf{R}}_{X+}(\tau+1)\hat{\mathbf{H}}(\tau)}{||\hat{\mathbf{H}}(\tau)||^2}. \tag{13}$$

The spectra of impulse responses $\hat{H}$ are blindly estimated using this VSS-UMCLMS.

3 Experiments

3.1 Evaluation data

A. Real reverberant speech with stationary noise

To evaluate our dereverberation method in a real environment with stationary noise, we recorded multi-channel speech degraded simultaneously by stationary noise and reverberation. Table 1 gives the conditions and content of the recordings. One hundred utterances from the Japanese Newspaper Article Sentences (JNAS) corpus, uttered by five male speakers seated on the chairs labeled A to E in Fig. 2, were recorded by a multi-channel recording device. The heights of the microphone array and the utterance position of each speaker were about 0.8 m and 1.0 m, respectively. An electric fan with high air volume is located behind the speaker in position A and was used as background noise. An average SNR of the speech was about 18 dB. We used a microphone array with 9 channels (Fig. 2) and a pin microphone to record speech in the distant-talking environment and close-talking environment, respectively.

B. Real reverberant speech with non-stationary noise

To evaluate our dereverberation method in a real environment with non-stationary noise, we recorded multi-channel speech degraded simultaneously by music and reverberation. One hundred utterances from the JNAS corpus, uttered by one male speaker seated on the chair labeled A in Fig. 2, were recorded by a multi-channel recording device. We played monaural music continuously until one hundred utterances was finished. Other recording conditions were same as real reverberant speech with stationary noise. An average SNR of the speech was about 3.4 dB.

Table 1: Conditions for recording.

microphone	SONY ECM-C10
A/D board	Tokyo Electron device TD-BD-16ADUSB
recording room size [m]	7.1(D) × 3.3(W) × 2.5(H)
number of speakers	5 male speakers
number of utterances	100 utterances (about 20 utterances per speaker)
sampling frequency	16 kHz
quantization bit rate	16 bits

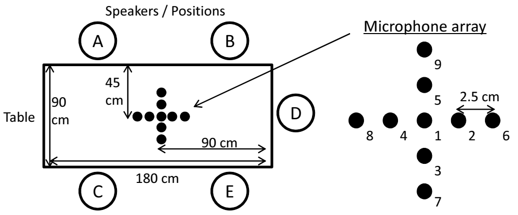


Fig 2: Illustration of recording settings and microphone array

3.2 Experimental results for reverberant speech with stationary noise

The speech recognition result is shown in Table. 2. To suppress the stationary noise, we proposed the combined use of dereverberation method with stationary noise reduction based on GSS. The schematic diagram of this method is shown in Fig. 3. For our dereverberation method, the parameters D, α, β, n were 6 (192 ms), 0.1, 0.15, and 0.1, respectively. We investigated the best channel combination in the real environment and the best speech recognition performance was obtained when channels 6, 7, 8, and 9 described in Fig. 2 were used. Therefore, this channel combination were used for blind estimation of impulse response by MCLMS and delay-and-sum beamforming. In Table. 2, "CMN only", "GSS" and "GSS+MCLMS" are the results which using the conventional CMN, stationary noise reduction based on GSS, combining GSS with our dereverberation method based on GSS by MCLMS, respectively. In this paper, delay-and-sum beamforming was performed for all methods.

In Table. 2, the word accuracy rate for close-talking speech recorded in a real environment was 88.3%. "GSS" method achieved a smaller improvement in recognition performance. On the other hand, speech recognition performance was significantly improved by "GSS+MCLMS" method compared to both the "CMN only" and "GSS" methods for almost all speakers. "GSS+MCLMS" method achieved an average relative word error reduction rate of 39.1% compared to "CMN only".

3.3 Experimental results for reverberant speech with non-stationary noise

The speech recognition result is shown in Table. 3. To suppress the non-stationary noise, we proposed the combined use of dereverberation method with blind source separation method based on EFICA [4] by T-ABCD tool [5]. The

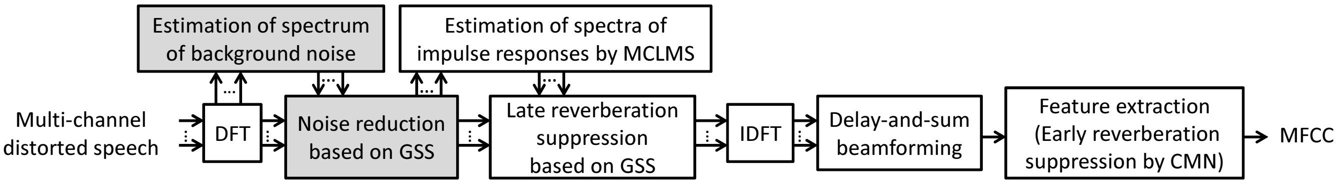

Estimation of spectrum of background noise
Estimation of spectra of impulse responses by MCLMS
Multi-channel distorted speech
DFT
Noise reduction based on GSS
Late reverberation suppression based on GSS
IDFT
Delay-and-sum beamforming
Feature extraction (Early reverberation suppression by CMN)
MFCC

Fig 3: Schematic diagram of our dereverberation method with stationary noise reduction

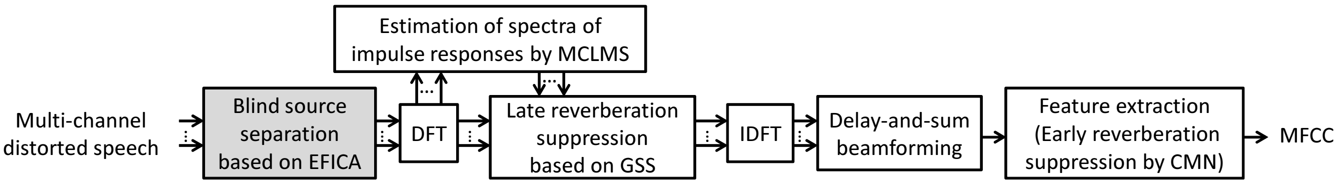

Estimation of spectra of impulse responses by MCLMS
Multi-channel distorted speech
Blind source separation based on EFICA
DFT
Late reverberation suppression based on GSS
IDFT
Delay-and-sum beamforming
Feature extraction (Early reverberation suppression by CMN)
MFCC

Fig 4: Schematic diagram of our dereverberation method with non-stationary noise reduction

Table 2: Word accuracy for LVCSR using the real reverberant speech with stationary noise (%).

Speakers / Position	CMN only	GSS	GSS+MCLMS
A	60.2	64.7	79.5
B	75.6	72.5	83.2
C	67.4	66.7	77.5
D	59.1	60.8	78.7
E	42.9	50.0	61.7
Average	60.9	62.9	76.2

Table 3: Word accuracy for LVCSR using the real reverberant speech with non-stationary noise (%).

Environments	CMN only	EFICA	EFICA+MCLMS
w/o music	46.1	57.4	70.2
with music	12.4	30.4	49.1

schematic diagram of this method is shown in Fig. 4. Four channels (Mic. 6, 7, 8 and 9 in Fig. 2) were used for blind estimation of impulse response by MCLMS, EFICA and delay-and-sum beamforming. In Fig. 3, "EFICA" and "EFICA+MCLMS" are the results which using blind source separation based on EFICA, combining EFICA with our dereverberation method based on GSS by MCLMS, respectively. The environment "w/o music" indicates the result of the speech withiout music.

In Table. 3, the speech recognition performance of "CMN only" was drastically degraded owing to the noisy reverberant conditions and the fact that CMN did not suppress the music and late reverberation. "EFICA" improved the speech recognition performance significantly compared to "CMN only". "EFICA+MCLMS" markedly improved compared to "EFICA", and outperformed than all the other methods. Our proposed method achieved an average relative word error reduction rate of 41.9% compared to "CMN".

4 Conclusion and Future Work

In this paper, we evaluated our dereverberation method based on GSS by MCLMS in real noisy reverberant environment with stationary noise and non-stationary noise. To suppress the stationary noise and non-stationary noise, we use a noise reduction technique based on GSS and a blind source separation based on EFICA, respectively. Under the stationary noise, our dereverberation method with GSS achieved an average relative word error reduction rate of 39.1% compared to conventional CMN. Under the non-stationary noise of a background music, our dereverberation method with EFICA achieved an average relative word error reduction rate of 41.9% compared to conventional CMN.

In the future, we intend to extend our proposed method to deal with real-world speech data including overlapping speech that involves multiple persons speaking simultaneously.

References

[1] L. Wang, N. Kitaoka and S. Nakagawa, "Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm, " IEICE Trans. Information and Systems, Vol.E94-D, No.3, pp. 659-667, Mar. 2011.

[2] Y. Huang, J. Benesty and J. Chen, "Acoustic MIMO Signal Processing," Springer, 2006.

[3] K. Odani, L. Wang, A. Kai, "Blind Dereverberation Based on Generalized Spectral Subtraction by Multi-channel LMS Algorithm, " Proc. of APSIPA ASC 2011, Oct. 2011.

[4] Z. Koldovský, P. Tichavský and E. Oja, "Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound," IEEE Trans. on Neural Networks, Vol. 17, No. 5, pp. 1265-1277, September 2006.

[5] Z. Koldovský and P. Tichavský, "Time-Domain Blind Separation of Audio Sources on the basis of a Complete ICA Decomposition of an Observation Space," IEEE Trans. on ASLP, Vol. 19, No. 2, pp. 406-416, Feburary 2011.

[6] K. Odani, L. Wang, A. Kai, "Speech Recognition by Blind Source Separation And Dereverberation Method for Mixed Sound of Speech And Music, " Proc. of ICASSP 2013 (to appear).

Speaker Verification under Channel Mismatch Condition

Ikuya Hirano * (hirano@spa.sys.eng.shizuoka.ac.jp)
Graduate School of Engineering, Shizuoka University, Japan
(Advisers: Longbiao Wang, † Atsuhiko Kai ‡ and Seiichi Nakagawa §)

1 Introduction

In conventional speaker verification methods based on mel-frequency cepstral coefficients (MFCCs), only the magnitude of the Fourier Transform in time-domain speech frames has been used. This means that the phase component has been ignored. Importance of phase in human speech recognition has been reported in [1], [2]. Several studies have invested great effort in modeling and incorporating the phase into the speaker recognition process [3].

Previously, Wang et al. proposed a speaker verification system using a combination of MFCCs and phase information [4], [5] directly extracted from the limited bandwidth of the Fourier transform of the speech wave. However, problems occurred in extracting the phase information because of the influence of the windowing position. Shimada et al. proposed a new method to extract pitch synchronous phase information [6]. The experimental results showed that the phase information was effective for speaker recognition under channel matched condition [4], [5], [6].

However, phase drastically changes between different channels. In [7], the experimental results indicated that the speaker recognition performance based on phase information was drastically degraded under channel mismatched and channel distortion conditions. To mitigate the influence of channel mismatch for phase information, joint factor analysis (JFA) [8] instead of traditional GMM-UBM based on Gaussian mixture model (GMM) and Universal background model (UBM) is used in this study. Recently, the JFA approach has become the active field for speaker verification. This modeling proposes powerful tools for addressing the problem of speaker and channel variability in GMM framework. Therefore, it is considered that the degradation of speaker verification performance using phase information under channel mismatched condition would be mitigated partly. Furthermore, a combination of the phase information-based JFA and MFCC-based JFA is also studied in this paper.

2 Phase Information Extraction

The spectrum $S(\omega, t)$ of a signal is obtained by DFT of an input speech signal sequence

$$\begin{aligned} S(\omega,t) &= X(\omega,t) + jY(\omega,t) \\ &= \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)}. \end{aligned} \qquad (1)$$

However, the phase $\theta(\omega, t)$ changes according to the frame position in the input speech. To overcome the influence of the phase response with respect to frame position, phases with the anchoring radian frequency ω_b for all frames are converted to a constant, and the phase with the other frequency is estimated relative to this. In the experiments discussed in this paper, the anchoring radian frequency ω_b is set to $2\pi \times 1000$ Hz. Actually, this constant phase value of the anchoring radian frequency does not affect the speaker recognition result. Without loss of generality, setting the phase with the anchoring radian frequency $\theta(\omega_b, t)$ to 0, we have

$$S'(\omega_b,t) = \sqrt{X^2(\omega_b,t)+Y^2(\omega_b,t)} \times e^{j\theta(\omega_b,t)} \times e^{j(-\theta(\omega_b,t))}, \qquad (2)$$

whereas for the other frequency $\omega = 2\pi f$, the spectrum on frequency ω is normalized as

$$\begin{aligned} S'(\omega,t) &= \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)} \times e^{j\frac{\omega}{\omega_b}(-\theta(\omega_b,t))} \\ &= \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\tilde{\theta}(\omega,t)} \\ &= \tilde{X}(\omega,t) + j\tilde{Y}(\omega,t). \end{aligned} \qquad (3)$$

Then, the real and imaginary parts of (3) are given by

$$\tilde{X}(\omega,t) = \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times \cos\left\{\theta(\omega,t) + \frac{\omega}{\omega_b}(-\theta(\omega_b,t))\right\}, \qquad (4)$$

$$\tilde{Y}(\omega,t) = \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times \sin\left\{\theta(\omega,t) + \frac{\omega}{\omega_b}(-\theta(\omega_b,t))\right\} \qquad (5)$$

and the phase information is normalized as

$$\tilde{\theta}(\omega,t) = \theta(\omega,t) + \frac{\omega}{\omega_b}(-\theta(\omega_b,t)). \qquad (6)$$

[*] He is studying in the 1st year at the course of Systems Engineering of this graduate school (Master course).

[†] He has been an associate professor at Nagaoka University of Technology.

[‡] He has been an associate professor in the Faculty of Engineering at Shizuoka University.

[§] He has been a professor at Toyohashi University of Technology.

In a previous study, to reduce the number of feature parameters, we used phase information in a sub-band frequency range only. However, a problem arose with this method when comparing two phase values. For example, for two values $\pi - \tilde{\theta}_1$ and $\tilde{\theta}_2 = -\pi + \tilde{\theta}_1$, the difference is $2\pi - 2\tilde{\theta}_1$. If $\tilde{\theta}_1 \approx 0$, then the difference $\approx 2\pi$, despite the two phases being very similar to each other. Therefore, we mapped the phase into coordinates on a unit circle [4], [5], that is,

$$\tilde{\theta} \rightarrow \{cos\tilde{\theta}, sin\tilde{\theta}\}. \tag{7}$$

Using the relative phase extraction method that normalizes the phase variation with respect to frame positions, the phase variation can be reduced. However, the normalization of phase variation is still inadequate. For example, for a 1000 Hz periodic wave (16 samples per cycle for a 16 kHz sampling frequency), if one sample point shifts in the cutting position (frame position), the phase shifts only $\frac{2\pi}{16}$, while for a 500 Hz periodic wave, the phase shifts only $\frac{2\pi}{32}$ with this single sample cutting shift. On the other hand, if the 17 sample points shift, their phases will shift by $\frac{17\cdot 2\pi}{16}(mod2\pi) = \frac{2\pi}{16}$ and $\frac{34\pi}{32}$, respectively, for the two periodic waves. Therefore, the values of the relative phase information for different cutting positions are very different from those of the original cutting position. We have addressed such variations using a statistical distribution model of GMM [4], [5].

If we could split the utterance by each pitch cycle, changes in the phase information would be further obviated. Thus, we propose a new extraction method that synchronizes the splitting section with a pseudo pitch cycle.

With respect to how to unite the cutting sections in the time domain, the proposed method looks for the maximum amplitude at the center around the conventional target splitting section of an utterance waveform, and the peak of the utterance waveform in this range is adopted as the center of the next window. Fig. 1 outlines how to synchronize the splitting section.

In this paper, however, we don't discuss the comparison with traditional phase information and pseudo-pitch synchronous phase information because the effectiveness of pseudo-pitch synchronous phase information has already been shown in [6].

3 Joint Factor Analysis

Joint factor analysis is an effective model for speaker verification under channel mismatched conditon. In this model, each speaker is represented by the means, covariance, and weights of a mixture of multivariate diagonal-covariance Gaussian densities defined in some continuous feature space of dimensions. The GMM for a target speaker is derived by adapting the universal background model (UBM) mean parameters. The basic assumption in JFA is shown as (8).

$$M = s + c, \tag{8}$$

where M is a speaker- and channel-dependent supervector, and s and c are speaker and channel supervectors, respectively.

The first term in the right-hand side of (8) is modeled by supposing that if s is the speaker supervector for a rondomly chosen speaker, then

$$s = m + Dz + Vy, \tag{9}$$

where m is the speaker- and channel-independent supervector (UBM), D is a diagonal matrix, V is a rectangular matrix of low rank, and y and z are independent random vectors which have standard normal distributions. The components of y and z are refered to the speaker and residual factors, respectively.

The channel-dependent supervector c, which represents channel effects in a speech, is supposed to be distributed according to

$$c = Ux, \tag{10}$$

where U is a rectangular matrix of low rank, and x has standard normal distribution. The components of x are refered to the channel factors.

For scoring, we used the linear scoring method. The equation of linear scoring method is shown as (11).

$$Score_{JFA} = (Vy + Dz)^{*} \times \Sigma^{-1} \times (F(test) - N(test)m - N(test)Ux), \tag{11}$$

where $Score_{JFA}$ is the speaker verification score based on JFA, Σ is a covariance matrix, and $N(test)$ and $F(test)$ are the zero and first order Baum-Welch statistics estimated from the feature of test utterances, respectively.

JFA realizes speaker verification that is robust for channel variabilities by two reasons, the division of the speaker model to two models that represent speaker and channel characteristics respectively for modeling and the calculation of the speaker verification score with removing channel variabilities for scoring.

A detailed description of JFA can be refered by literature [8].

4 Combination Method and Decision Method

MFCCs use only the magnitude of the Fourier Transform in time-domain speech frames, that is, phase component is ignored. On the other hand, phase information ignores the magnitude of the Fourier Transform in time-domain speech frames. Therefore, in this paper, the JFA score based on MFCCs is combined with the JFA score based on phase

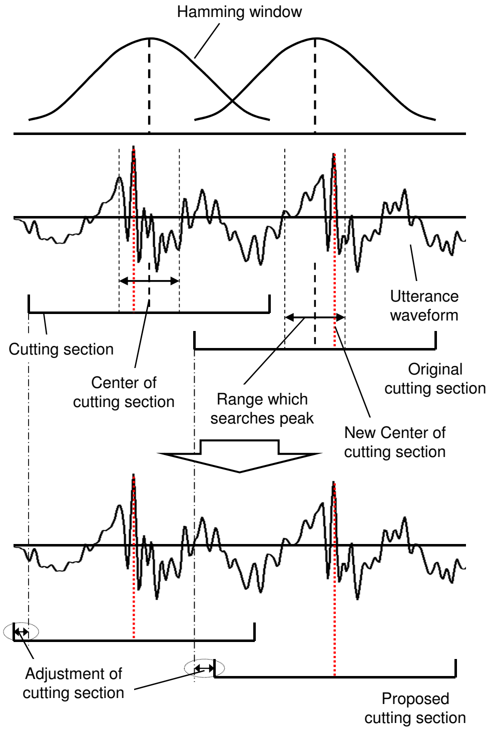

Hamming window
Utterance waveform
Cutting section
Center of cutting section
Range which searches peak
Original cutting section
New Center of cutting section
Adjustment of cutting section
Proposed cutting section

Fig. 1: How to synchronize the splitting section

Table. 1: Conditions for speech analysis

sampling frequency	8 kHz	
	MFCC	phase
window size	25 ms	16 ms
window shift	10 ms	5 ms
frequency range	all	60-700 Hz
dimensions	60 (19 MFCCs + power, their Δ and ΔΔ coefficients	24 (12 sin and 12 cos components

Table. 2: EERs for speaker verification (%)

		GMM-UBM	JFA
MFCC	male	7.71	6.97
	female	7.38	3.44
phase	male	24.57	19.43
	female	17.86	14.76
MFCC+phase	male	7.68	6.72
	female	6.65	3.25

information. When a combination of the two methods is used to identify the speaker, the score of the MFCC-based JFA is linearly coupled with that of the JFA based on phase information to produce a new score $Score_{comb}$ given by

$$Score_{comb} = (1-\alpha)Score_{MFCC} + \alpha Score_{phase}, \tag{12}$$

where $Score_{MFCC}$ and $Score_{phase}$ are the score produced by MFCC-based speaker model and phase information-based speaker model, respectively, and α denotes the weighting coefficients, which are determined empirically. The combination score is then compared to the threshold in order to take the final decision.

5 Experiments

5.1 Experimental setup

The effect of phase information-based JFA for speaker verification under channel mismatched condition was evaluated on the NIST 2003 SRE database [9]. The NIST 2003 SRE database consists of recordings of 356 speakers (149 males and 207 females), recorded in multiple conditions which include six transmission methods (CDMA, LAND, GSM, TDMA, CELLULAR and UNK), multiple telephones, multiple places, etc. Almost all the data for every speaker were recorded by different environments. Therefore, this speaker verification task is very difficult. We used gender-dependent UBMs containing 1024 Gaussians for MFCC and 256 Gaussians for phase information, respectively.

To verify robustness of phase information-based JFA for channel variability, the speaker verification system using JFA is compared with the system using traditional GMM-UBM in this paper. For GMM-UBM, GMMs containing 1024 Gaussians for MFCC and 256 Gaussians for phase information applying Maximum a posteriori (MAP) adaptation from gender-dependent UBMs were used. Table 1 shows conditions for the speech analysis.

5.2 Experimental results

The equal error rates (EERs) for speaker verification are given in Table 2. Phase information-based JFA showed the improvement of EERs of 5.14% for male and 3.10% for female compared to the system based on GMM-UBM. The results show that phase information-based JFA has the moderate performance for speaker verification and the degradation of speaker verification performance using phase information caused by speaker and channel variability was mitigated partly. The combination of MFCC and phase information achieved a better result than MFCC-based JFA which was one of the standard methods for speaker verification. This indicated that the phase information has complementary nature with MFCC.

These results show that phase information is effective for the speaker verification, even under channel mismatched condition. On the other hand, a previous study showed that phase information was not effective under channel mismatched and channel distortion conditions [7]. The reason is that the influence of channel mismatch for phase information is mitigated by using the channel variability robust JFA method. To verify this, we evaluated EERs for speaker verification under transmission mode matched and mismatched condition, respectively. We think that channel

Table. 3: EERs for speaker verification under transmission mode matched and mismatched condition (%)

		match		mismatch	
		GMM-UBM	JFA	GMM-UBM	JFA
MFCC	male	5.76	5.88	15.68	10.72
	female	6.21	3.17	13.46	5.73
phase	male	21.11	16.18	41.45	37.07
	female	11.22	10.91	42.06	34.48
MFCC+ phase	male	5.33	5.00	15.73	10.83
	female	5.25	3.20	12.50	5.72

characteristics varies drastically between different transmission modes. Here, transmission mode matched condition means that the enrollment utterance and test utterance have same transmission mode, while transmission mode mismatched condition means that the enrollment utterance and test utterance have different transmission mode. The experimental result is given in Table 3. For both MFCC and phase information, the system based on JFA showed the improvement of EERs compared to the system based on GMM-UBM under transmission mode mismatched condition. From Table 3, it is obvious that the JFA can partly remove the influence of transmission mode mismatch for phase information, but the influence is still large. Based on these results, the more improvement of the results shown in Table 2 are expected by normalizing phase information for each transmission mode.

6 Conclusions and Future Work

In this paper, we conducted the speaker verification using pseudo-pitch synchronous phase information under channel mismatched condition. To mitigate the influence of phase information under channel mismatched condition, a channel variability robust speaker verification method was applied. Phase information-based JFA showed the improvement of EERs of 5.14% for male and 3.10% for female compared to the traditional system based on GMM-UBM. We obtained the better result than only MFCC by combining MFCC and phase information.

Phase information shows the lower EERs under transmission mode matched condition while the higher EERs under transmission mode mismatched condition. As a result, in future work, we will try to normalize phase information for each transmission mode.

References

[1] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," *Proc. Eurospeech'03*, 2117-2120, 2003.

[2] G. Shi et al., "On the importance of phase in human speech recognition," *IEEE Trans. Audio, Speech, Lang. Process*, Vol.14, No.5, pp.1867-1874, Sep 2006.

[3] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker verification," *IEEE Signal Processing Letters*, Vol.13, No.1, pp.52-55, 2006.

[4] L. Wang, S. Ohtsuka and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," *Proc. ICASSP*, pp.4529-4532, 2009.

[5] S. Nakagawa, L. Wang and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech and Language Processing*, Vol.20, No.4, pp.1085-1095, 2012.

[6] K. Shimada, K. Yamamoto and S. Nakagawa, "Speaker identification using pseudo pitch synchronized phase information in voiced sound," *Proc. on APSIPA ASC 2011*, CD-ROM, Xi'an, China, Oct 2011.

[7] L. Wang and S. Nakagawa, "Speaker identification/verification for reverberant speech using phase information," *Proc. of WESPAC 2009*, No.0130 (8pages), 2009.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Trans. on Audio, Speech and Language*, Vol.16, No.5, pp.980-988, July 2008.

[9] "http://www.nist.gov/speech/tests/spk/2003/index.htm".

Singer recognition with accompaniment sound reduction and additional speaker feature through the use of phase information and separation of accompaniment sound and singing voice using the NMF

Ryouhei Nakaoka * (nakaoka@spa.sys.eng.shizuoka.ac.jp)
Faculty of Engineering, Shizuoka University, Japan
(Advisers: Longbiao Wang † and Atsuhiko Kai ‡)

1 Abstract

This paper describes the outline of our new research theme on musical information processing. We focus on the identification of the singers' name of the acoustic signal of music data. The study just started this year for my graduate work and some preliminary results based on a previous study[1] are reported. Remaining issues and future plans for this study are discussed.

2 Introduction

Information about the singers' name of the song is useful for music information retrieval (MIR). For example, if the name of a singer can be identified without any information of the meta-data about songs, users can find songs sung by a certain singer using a description of singers' names. Most MIR systems are based on meta-data and assume that the meta-data including artists' names and song titles are available: if they were not available for some songs, these songs could not be retrieved by submitting a query of their artists' names. Furthermore, detailed descriptions of the acoustic characteristics of singing voices can also play an important role in MIR systems because they are useful for systems based on vocal timbre similarity by computing acoustical similarities between singers. Hence, a user can discover new songs rendered by the singing voices they prefer.

In order to realize a music search based on the singers' name, the problem of identifying the singers' name of a song from musical audio signals is addressed. The main problem with the identification of singers' name is that a speaker specific characteristic is influenced by accompaniment sound as well as variations in acoustic features of the singer's voice. Therefore, it is difficult to accurately identify the singer without coping with such problems.

Previous study by Fujihara et al.[1] has proposed two methods to solve the problem of mixed sound accompaniment. They are "accompaniment sound reduction" and "reliable frame selection". Using the former method, we can reduce the influence of instrumental accompaniment. First, the harmonic structure of the melody is extracted from audio signals, and re-synthesize the melody by using a sinusoidal model. The latter method is used to select reliable frames that represent the characteristics of the singing voice. Fig.1 shows an overview of this method.

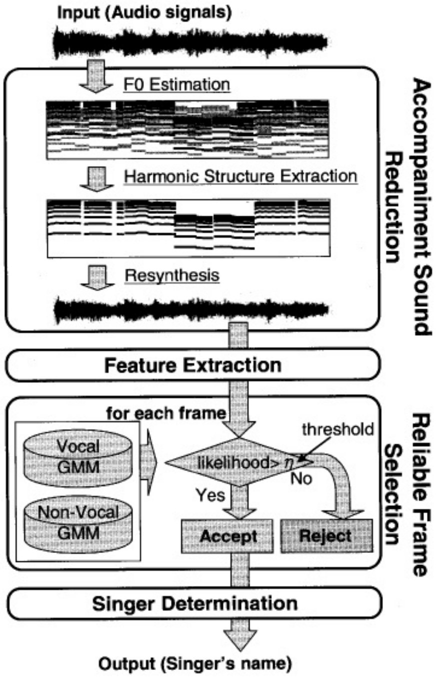


Fig.1. Overview of the baseline method[1].

3 Singer identification

Our approach of the singer identification system will be described.

[*] He is studying in the 4th year at the course of Systems Engineering of this university.
[†] He has been an associate professor at Nagaoka University of Technology.
[‡] He has been an associate professor in the Faculty of Engineering at Shizuoka University.

3.1 Accompaniment sound reduction

Previous study[1] proposed an accompaniment sound reduction method which is based on the extraction of the harmonic structure of the melody. However, this method relies on a F_0 estimation processing and it influences the accuracy of this step. Therefore, we plan to adopt a different approach which doesn't rely on an explicit F_0 estimation processing. Recently, non-negative matrix factorization (NMF), one of multivariate analysis methods is employed for many signal processing applications. Nakano et al.[2] at Toyohashi University of Technology proposed a method for speech recognition in mixed sound of speech and music based on a NMF approach. We are undertaking a joint study with Toyohashi University of technology and implementing. The accompaniment sound reduction by applying this NMF-based method.

3.2 Feature extraction

Detailed investigation of effective acoustic features for singer identification have not been performed so far. In the previous study[1], experiments were conducted to compare the performance between two types of acoustic feature parameters Mel Frequency Cepstral Coefficient(MFCCs) and Linear Prediction Mel Cepstral Coefficient(LPMCCs). Previous work[3] has shown that an additional speaker feature "phase informations" improves the performance of speaker identification and verification tasks. Thus, in order to express the individuality of the singer more robust, we will investigate to incorporate the phase information and explore the optimal combination of phase information, LPMCC and MFCC.

3.3 Reliable frame selection

Previous study[1] showed that a reliable frame selection is effective and we investigate to apply this method in addition to the above mentioned methods. This step uses a vocal GMM λ_V and a non-vocal GMM λ_N. The vocal GMM λ_V is trained on feature vectors extracted from the singing sections, and the non-vocal GMM λ_N is trained on those extracted from the interlude sections. Given a feature vector $\boldsymbol{x}$, the likelihoods for two GMMs, $p(\boldsymbol{x}|\lambda_V)$ and $p(\boldsymbol{x}|\lambda_N)$, that correspond to how likely the feature vector $\boldsymbol{x}$ is a vocal part or a (non-vocal) instrument part, respectively is computed. Therefore whether the feature vector $\boldsymbol{x}$ is reliable or not is determined by using the following equation:

$$\log p\left(\boldsymbol{x}|\lambda_V\right) - \log p\left(\boldsymbol{x}|\lambda_N\right) \underset{not-reliable}{\overset{reliable}{\gtrless}} \eta \tag{1}$$

where η is a threshold.

It is difficult to determine a universal constant threshold for a variety of songs because if the threshold is too high for some songs, there are too few reliable frames to appropriately calculate the similarities. Therefore, the threshold that is dependent on songs is determined so that the α% of all the frames in each song is selected as reliable frames. It is expected that most of the non-vocal frames are rejected in this selection step.

3.4 Determining the singers' name

For each singer to be identified, GMMλ_s(s is a singer's label) will be trained in advance. $\boldsymbol{X} = \{\boldsymbol{x_t}|t = 1, ..., T\}$ is a feature vector sequence chosen by the reliable frame selection. A singer's name is determined based on the following equation.

$$s = \arg\max_i \frac{1}{T}\sum_{t=1}^{T}\log p\left(\boldsymbol{x_t}|\lambda_i\right) \tag{2}$$

4 Experiments

4.1 Experimental condition

So far, we have done a baseline experiment which does not employs both of two methods, which are "reliable frame selection" and "accompaniment sound reduction.", described in section 3.1 and 3.3. Experiments of singer identification were conducted using the RWC Music Database: Popular Music (RWC-MDB-P-2001)[4]. 40 songs by ten different singers (five were males and five were females) taken from the RWC-MDB-P-2001 were used[1]. The data include four tracks per singer. Using these data, a four-fold cross validation was conducted. Three songs out of four songs are used for training singer models for each singer, and the remaining song is used in the evaluation. The number of mixtures for each singer's GMM is 64. The covariance matrix of Gaussian densities assumes diagonal covariance.

Table 1 shows the conditions of feature extraction stage.

[1] This condition is identical to the previous work[1].

TABLE.1. Feature extraction conditions.

#Channels	1
Sampling frequency	16kHz
Frame shift	10ms
Frame length	128ms
Window function	Hamming
Preemphasis coef	0.97
Feature vector	MFCC,LPMCC
Dimension	12(static coef only), 25(static coef + Δ + Δenergy)

4.2 Result

The results of singer identification experiments are shown in Figure 2. The accuracy was defined by a ratio of the number of correctly identified song to the number of songs used for evaluation.

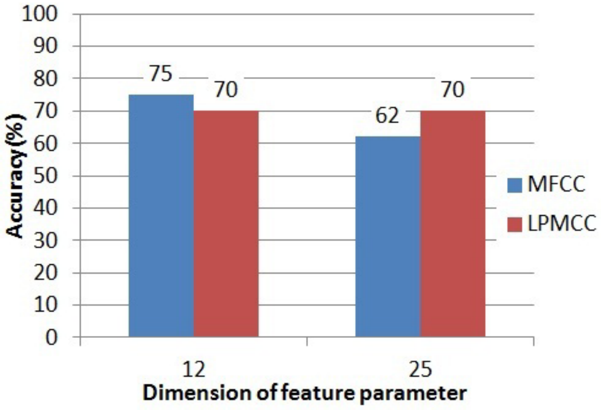


Fig.2. Results of singer identification experiments.

The best result is attained by using the 12th-order MFCC feature parameter. In the previous study[1], the results of a similar experiment using 12th-order MFCC and 12th-order LPMCC are shown. The accuracies were 53% for MFCC and 55% for LPMCC, respectively. At the time of writing, it is not clear why our experiments attained better results than previous study.

5 Progress situation and future work

Currently, we are in the process of doing the experiment of introducing accompaniment sound reduction and reliable frame selection described in section 3.1 and 3.3. Our future plan is as follows:

- Implement the baseline methods of accompaniment sound reduction(section 3.1) and reliable frame selection(section 3.3).
- Compare approaches for accompaniment sound reduction: baseline method[1] v.s. NMF based method[2].

We are now trying to investigate the effectiveness of a new method which incorporates two novel methods proposed in[2][3] to the singer identification framework described in section3. The accompaniment sound reduction method proposed in previous study[1] is based on an explicit F0 extraction and therefore the effectiveness of this method depends on the accuracy of the F0 extraction. On the other hand, F0-independent method[2]. In addition, we propose a method to combine the phase information to the feature extraction(section3.2). We will continue to advance the experiment to verify the usefulness of these methods.

Both of the previous work[1] and our experiment has used only 10 singers. We aim to be able to identify singers with high accuracy even for further increase a number of singers.

References

[1] H.Fujihara et al., “A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval,” IEEE Transactions on Audio, Speech, and Language Processing, Vol.18, No.3, pp.638-648, 2010.

[2] S.Nakano, K.Yamamoto and S.Nakagawa, “Speech recognition in mixed sound of speech and music based on vector quantization and non-negative matrix factorization,” Proc. of InterSpeech, pp.1781-1784 (August 2011)

[3] S.Nakagawa, L.Wang and S.Ohtsuka, “Speaker Identification and Verification by Combining MFCC and Phase Information,” IEEE Transactions on Audio, Speech, and Language Processing, Vol.20, No.4, pp.1085-1095, 2012.

[4] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “ RWCmusic database: Popular, classical, and jazz music databases, ” in Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR 2002), Oct. 2002, pp. 287.288.

Applying Deep Belief Network for Voice Activity Detection Task

Akihiro Nakatani * (nakatani@spa.sys.eng.shizuoka.ac.jp)
Faculty of Engineering, Shizuoka University, Japan
(Advisers: Longbiao Wang † and Atsuhiko Kai ‡)

1 Introduction

Recently, we started to explore the applications of deep belief network (DBN) on two speech-related subjects; speaker recognition task and voice activity detection (VAD) task. This paper describes the outline of deep belief network (DBN) and a plan of application to the voice activity detection task in adverse environment for my graduate work. Since my graduate work just started and related works have been surveyed, this paper introduces the main points of a most related work by A. Mohamed et al.[1]. After that, a research plan for my graduate work is described.

2 Acoustic Modeling Using Deep Belief Networks[1] : Survey

Gaussian mixture models (GMM) are currently the dominant technique for modeling the emission distribution of hidden Markov models (HMM) for speech recognition. In [1], it is shown that better phone recognition can be achieved by replacing Gaussian mixture models by deep neural networks that contain many layers of features and a very large number of parameters. It is shown that multilayer neural network is affected better by pre-training .

2.1 Learning the ganerative multi-layer model

(1) Restricted Boltzmann Machine

In order to learn the generative multi-layer model, consider the first two layers of binary RBM. Fig.1 shows that Restricted Boltzmann Machine is a model represented by the input unit representing the observed values, hidden units representing the features are bound to each other, as the two portions of the undirected graph.

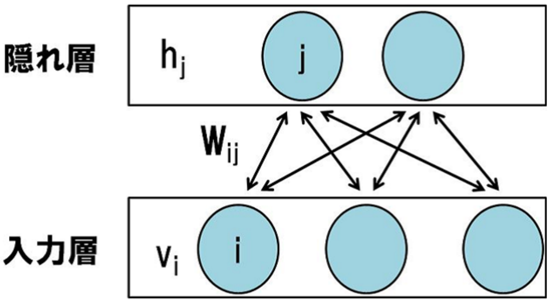


Fig1. RBM configuration

The binary RBM, via the energy function, the weight of the individual units connected to the bias defines a probability distribution for each state of the input unit and the hidden units. Set of energy for each state are shown as follows.

$$E(\mathbf{v},\mathbf{h}|\theta) = -\sum_{i=1}^{\mathcal{V}}\sum_{j=1}^{\mathcal{H}} w_{ij}v_ih_j - \sum_{i=1}^{\mathcal{V}} b_iv_i - \sum_{j=1}^{\mathcal{H}} a_jh_j \tag{1}$$

b_i and a_j is a bias term,、 $\theta = (\mathbf{w},\mathbf{b},\mathbf{a})$ and ω_{ij} denotes the symmetric interaction term between input unit i and hidden unit j. $\mathcal{V}\ and\ \mathcal{H}$ are the number of visible and hidden units. The probability that an RBM assigns to a visible vector $\mathbf{v}$ is

$$p(\mathbf{v}|\theta) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{u}}\sum_{\mathbf{h}} e^{-E(\mathbf{u},\mathbf{h})}} \tag{2}$$

Since there are no hidden-hidden connections, the conditional distribution $p(\mathbf{h}|\mathbf{v},\theta)$ is factorial and is given by

$$p(h_j = 1|\mathbf{v},\theta) = \sigma\left(a_j + \sum_{i=1}^{\mathcal{V}} w_{ij}v_i\right) \tag{3}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$. Similarly, since there are no visible-visible connections, the conditional distribution is factorial and is given by

$$p(v_j = 1|\mathbf{h},\theta) = \sigma\left(b_i + \sum_{j=1}^{\mathcal{H}} w_{ij}h_j\right) \tag{4}$$

Exact maximum-likelihood learning is infeasible in a large RBM because it is exponentially expensive to compute the derivative of the log probability of the training data. However, we will be able to learn effectively by learning

[*] He is studying in the 4th year at the course of Systems Engineering of this graduate school (Bachelor course).
[†] He has been an associate professor at Nagaoka University of Technology.
[‡] He has been an associate professor in the Faculty of Engineering at Shizuoka University.

approximation called ”contrastive divergence”. We repeatedly update each weight w_{ij} using the difference between two measured, pairwise correlations

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{reconstruction} \tag{5}$$

The first term is the expected value of the frequency in which the unit i and hidden unit j input are on together. The second term is the measured frequency with which and are both on when the visible vectors are ”reconstructions” of the data vectors and the states of the hidden units are determined by applying (3) to the reconstructions.

(2)Reconstructions

This section describes the details on how to calculate the reconstructions that were introduced in the previous section. First, the input data is applied to the input layer, and the values of hidden layer are calculated. ”Reconstructins” is a calculation method to determine the values of the input layer again with the values of the hidden layer. We proceed to update the weights using Equation (5) and this reconstruction.

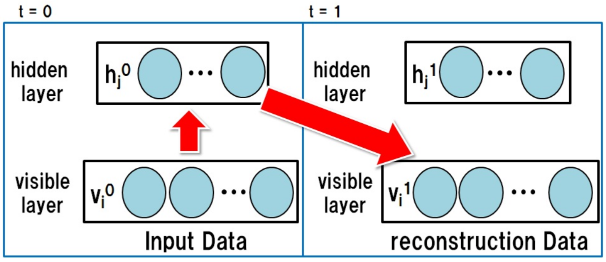


Fig.2. how to calculate the reconstructions

(3)Construction of DBN

In this way, we can learn a hierarchical model stacked by RBM. A multilayer generative model that is pre-trained, called Deep Belief Net (DBN), is constructed. At that time, throw away all the layers except the top layer, add a layer of the label ”softmax” unit that represents the state of the HMM. Further, when dealing with the entire system as a feed-forward neural network, this network is discriminatively fine-tuned by using backpropagation to maximize the log probability of the correct HMM state.

2.2 Using DBN for phone recognition

In order to apply DBNs with fixed input and output dimensionality to phone recognition, we use a context window of successive frames of speech coefficients to set the status of the visible units of the lowest layer of the DBN. Feed-forward neural networks are learned so as to output a probability distribution that was estimated for the label of the central frame. To generate phone sequences, the sequence of predicted probability distributions over the possible labels for each frame is fed into a standard Viterbi decoder.

2.3 Experiments

Fig. 3 shows the results of two compared systems which was pre-trained (pretrain) or initiated from a random weight (rand) respectively. A single hidden layer 2048 is almost no difference between pretrain and rand as shown in Fig.3. Therefore, generative pre-training will not be affected much in a single hidden layer . However, the difference between the two (rand and pretrain)is noticeable about increasing the number of layers. Thus, we can obtain the benefits of prior learning by increasing the number of layers. Compared to the method described in this paper (Monophone DBN) and the existing approaches (CD-HMM), PER (phone error rate) has been improved by approximately 7 %.

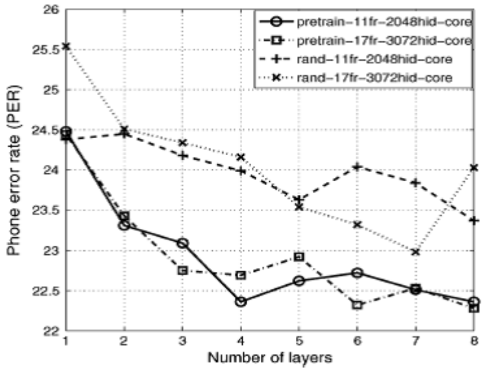


Fig.3 Phone error rate on the core test set as a function of the number of hidden layers using randomly initialized and pretrained networks.[1]

3 Applying DBN for VAD task

3.1 Our approach

We started to study an VAD system which identifies of the speech segment and non-speech segments based on the DBN model. In related works on large vocabulary speech recognition systems, the DBN exhibits the high accuracy and robustness to various speech contents. Thus it is expected that the VAD system based on the DBN outperforms the existing VAD approaches, when they applied to degraded speech in adverse environment. To train and test the DBN model, it is necessary to label the speech data. The frame-level label of speech/non-speech given to the neural network in correspond to the input feature vector, MFCC.

3.2 Voice activity detection using MLP

Before pursing with the goal of the VAD task with DBN, a VAD method based on multilayer perceptron (MLP), is investigated as a baseline result. MLP is composed of more than one hidden layer, input layer, and output layer. One or more units present in each layer of the MLP, which are connected in one unit present in the surrounding layers. MLP used in this study is a neural network with one hidden layer as shown in Fig. 4. The main difference between MLP and DBN is that the latter employs RBM for pretraining, though the former does not employ it. We will compare the result in the same way as we described in the section 2.3.

3.3 Current progress and Future work

We intend to implement VAD using the aurora-front-end that is VAD algorithm which has the MLP[2]. Japanese Electronic Industry Development Association (JEIDA) corpus is used as speech data.
Future work are the following three

- Train MLP as speech/non-speech classifier(baseline method)
- Train DBN-based speech/non-speech classifier
- Compare two approaches by evaluation experiments

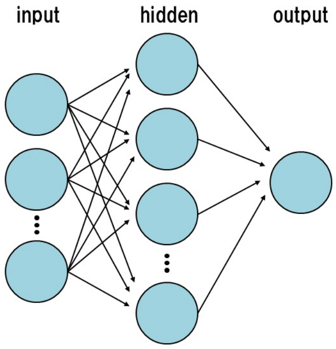


Fig.4 A single hidden layer MLP

References

[1] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton, "Acoustic Modeling Using Deep Belief Networks" IEEE Transactions on Audio, Speech, and Language Processing, Vol.20, pp.14-22 2012

[2] E. Zwyssig, S. Renals, and M. Lincoln. "Determining the number of speakers in a meeting using microphone array features. "In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4765-4768, 2012

NANYANG TECHNOLOGICAL UNIVERSITY

Collection and Annotation of Malay Conversational Speech Corpus

presented by Chng Eng Siong

Tze Yuang Chong[1,3], Xiong Xiao[1], Tien-Ping Tan[2], Eng Siong Chng[1,3], Haizhou Li[1,3,4]

[1]Temasek Laboratory, Nanyang Technological University, Singapore

[2]School of Computer Science, Universiti Sains Malaysia, Malaysia

[3]School of Computer Engineering, Nanyang Technological University, Singapore

[4]Institute for Infocomm Research, Singapore

11 December 2012

Outline

- Motivation
- Objectives
- Specification
- Recording Setup
- Transcription & Annotation
- Demographic
- Analysis

NANYANG TECHNOLOGICAL UNIVERSITY

Motivation

- Conversational LVCSR system
  - To extend our effort from broadcast news and read speech recognition to conversational speech recognition
  - With our existing skills on planned speech recognition, explore new challenges on spontaneous speech recognition

- Malay as a local language
  - Less attention is paid to Malay LVCSR yet it is the major language spoken by 14 million people
  - No Malay conversational speech corpus is distributed so far – sparse resource

Objectives

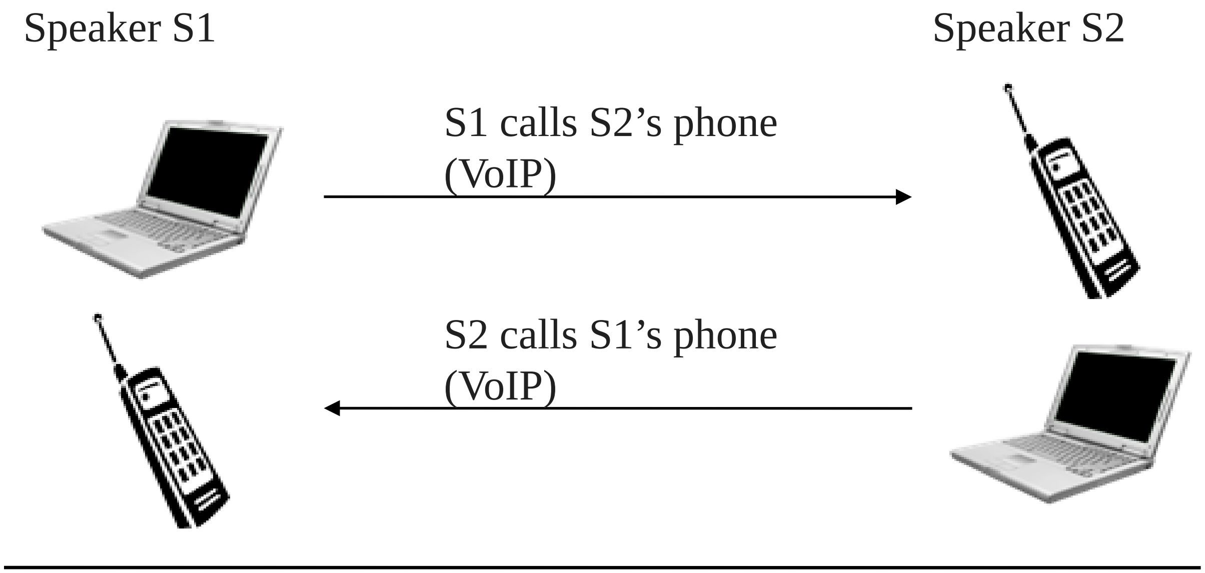
- Record and transcribe 50 hours of spontaneous conversational speech of Malay language
  - To provide training and evaluation material for the development of the Malay LVCSR system
  - To support the research of spontaneous speech recognition
    - Acoustic modeling – variation of speaker pronunciation, channel difference, emotion state, etc.
    - Language modeling – ungrammatical and fragmented sentence, disfluencies, etc.

Specifications

- 50 hours of spontaneous conversational speech in Malay language
  - Capture the speaker variation
    - More speakers for participation
    - Gender balance
    - Robustness
  - Expand the vocabulary
    - Prepare a topic pool for conversation
    - Collect more words that relate to a topic
  - Diversify the recording channel
    - Record speech from both the close-talk channel (clean speech) and telephone channel (noisy speech) simultaneously
    - Every utterance will be recorded in two conditions
  - Metadata tagging
    - Some metadata will be selected for annotation – to enhance the readability of the transcription

Recording Setup

- Record the clean speech and the telephone speech simultaneously from an emulated telephone conversation
  - Bidirectional communication
  - Collect speech from two conditions/channels



Speaker S1

Speaker S2

S1 calls S2's phone (VoIP)

S2 calls S1's phone (VoIP)

Recording Output:

S1's clean speech
S2's telephone speech

S2's clean speech
S1's telephone speech

Recording Setup

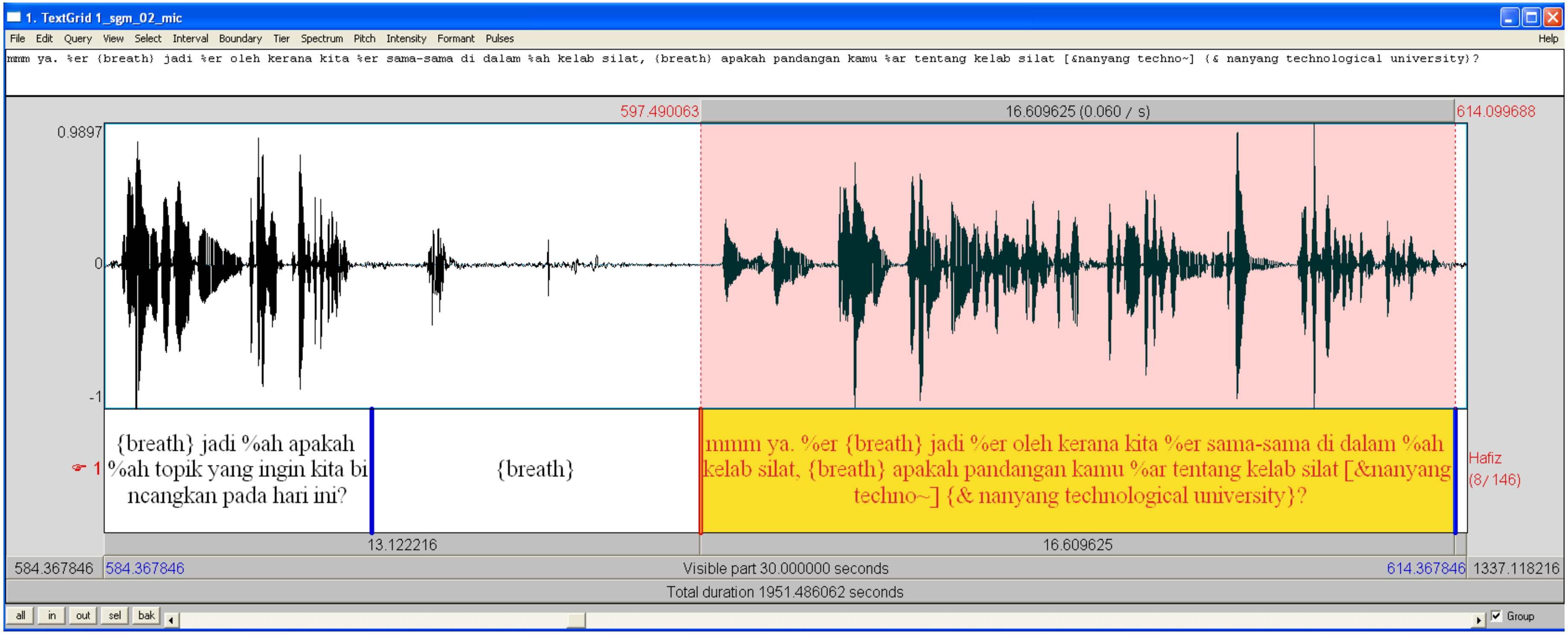
- Requirement:
  - 2 PCs, installed with
    - VoIP software and recording software
    - Soundcards that contain two recording devices (to record both the close-talk speech and the telephone speech simultaneously)
  - 2 phones
- Procedure:
  - To establish the communication, S1 and S2 use the VoIP software (on the PC) to dial each other's phone
    - S1's speech is sent through the telephone network and listened by S2
    - S2's speech is sent through the telephone network and listened by S1
  - Speech from both channels (close-talk and telephone) can be recorded simultaneously
    - S1's PC records S1's close-talk speech (from the microphone) and S2's telephone speech (from the VoIP software)
    - S2's PC records S2's close-talk speech (from the microphone) and S1's telephone speech (from the VoIP software)

Transcription and Annotation

- Word transcription
  - Verbatim
  - Utterances are time segmented based on sentence/sentence-unit (semantically complete segment)
- Metadata annotation
  - Linguistic events:
    - **Disfluencies** – The unsmooth or incomplete segment
      - E.g. “This is a cat, I mean, a dog.”
    - **Filled pauses** – The hesitation sounds during speaking
      - E.g. “This is uh a dog.”
    - **Partial words** – The incomplete words pronounced
    - **Code-switched words** – The English (or other languages) words mixed with Malay utterances
  - Audio events:
    - **Speaker’s noise**, such as cough, breath, lip smack, laugh, sniff, etc.
    - **Environmental noise**, such as flipping papers, clicking mouse, knocking table, etc

Transcription and Annotation

- Transcription with Praat



```
mmm ya. %er {breath} jadi %er oleh kerana kita %er sama-sama di dalam %ah kelab silat, {breath} apakah pandangan kamu %ar tentang kelab silat [&nanyang techno~] {& nanyang technological university}?
```

{breath} jadi %ah apakah %ah topik yang ingin kita bi ncangkan pada hari ini?

{breath}

mmm ya. %er {breath} jadi %er oleh kerana kita %er sama-sama di dalam %ah kelab silat, {breath} apakah pandangan kamu %ar tentang kelab silat [&nanyang techno~] {& nanyang technological university}?

Demographic

	Singapore	Malaysia	Total	Remarks
Sessions	39	81	120	
Duration	16 hours	34 hours	50 hours	Ave. 25min/session
Speakers (F/M)	45 (26/19)	54 (26/28)	99 (52/47)	Each speaker averagely contributes 20-40 min
Topics	18	41	59	Ranges from *hobby*, *travelling* to *social issues* and *politics*. About 50min/topic
Words	141K	294K	435K	Speaking rate is about 150 words/min in general

NANYANG TECHNOLOGICAL UNIVERSITY

Corpus Analysis

- Sentence segmentation
  - Utterances have been time-segmented to mark the boundaries of the sentences or the semantically complete sentence-unit.
    - A complete sentence

      "*dia orang daripada pagi sampai malam judi tak habis habis*."
      "*they gamble from day to night all the time."* (in English)
    - Back-channel (i.e. the respond to the active speaker)

      "*hmm*"
      "*uh-huh*"
    - Short sentence-unit

      "walaupun saya suka barang tu, hmm."
      "althought I love that thing, hmm." (in English)
  - The sentence/sentence-unit boundaries serve the ground truth for the task of sentence boundary detection
    - Does not fully support SU subtype detection task (i.e. statement, backchannel, question, incomplete) at the moment

NANYANG TECHNOLOGICAL UNIVERSITY

Corpus Analysis

- Filled pauses and disfluencies
  - As the conversation is spontaneous, utterances are common unsmooth and contain filled pauses and (edit) disfluencies.
  - For edit disfluency, we followed the LDC's SimpleMDE specification, the reparandum (deletable region) is tagged.
    - For example,

“*pasal kadang-kadang* [***tak boleh***] *tak boleh salahkan dia orang kalau dia orang masuk ini semua,* [***memang***] *memang bukan salah dia orang ah*.”

“*sometimes (we)* [***cannot***] *cannot simply blame them for joining (the triad),* [***actually***] *actually it's not all their fault*.” (in English)

“*aktiviti di mana* [***mereka boleh bergaul***] *%uh pekerja asing ini boleh bergaul dengan %um orang awam*.”

“*activitity where* [***they could socialize***] *%uh these foreign workers could socialize with %um the public*.”

NANYANG TECHNOLOGICAL UNIVERSITY

Corpus Analysis

- Filled pauses and disfluencies
  - Filled pause are labeled to a filler word that best pronounce the sound.
    - E.g., “*err*”, “*um*”, “*uh*”, etc.
    - They will be tagged (e.g. preceded with “%” sign) to be distinguished from other lexemes.
    - For example,

“ini kerana ***%um*** *penduduk* ***%uh*** *akan terasa terganggu.”*

“this is because ***%um*** *the residents* ***%uh*** *will find themselves being disturbed.*” (in English)

“***%uh*** *saya melakukan diploma di politeknik.* ***%um*** *peperiksaan atau* ***%um*** *penilaiannya tidak bergantung pada satu peperiksaan saja*.”

“***%uh*** *I obtained my diploma from polytechnic.* ***%um*** *The examination or* ***%um*** *evaluation didn’t depend on only one examination*.” (in English)

NANYANG
TECHNOLOGICAL
UNIVERSITY

Corpus Analysis

- Filled pauses and disfluencies
  - Edit words account 3.7% of the total words in the corpus.
  - Filler word is the most frequent token that accounts 5.6% of the total words in the corpus

  - Supporting the research of edit disfluency detection and filler detection
    - Enhance the readability of the transcription
    - Recognition output could be used for higher level natural language processing task, such as understanding and translation

Corpus Analysis

- Code-switching
  - It is very common mixing English words into Malay conversation, especially in the informal talking.
  - For example,

    "***$but $I $mean $maybe $his $luck*** *ah eh. Nasib juga lah, nasib sendiri*."

    "*but I mean maybe that was his lucky. It's fate, his own fate.*" (in English)

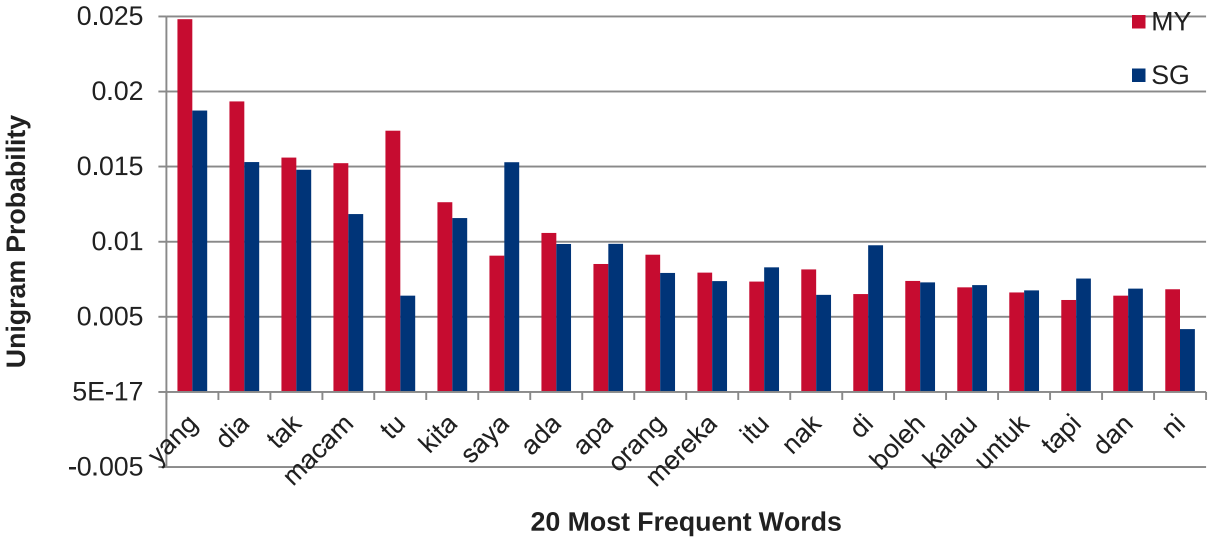
    "*pada pandangan anda, sudah mencukupi* ***$or*** *terlalu banyak ke* ***$or*** *macam sikit ke*?"

    "*in your opinion, is it adequate or it's excessive or too little*?" (in English)

  - English words account 2.8% of the total words in the corpus and 15.7% of the fallback vocabulary
  - Some most common English words in the corpus are "*so*", "*I*", "*you*", "*then*", "*time*", "*best*", etc.

Corpus Analysis

- Sociolinguistic difference
  - – The following figure shows the unigram probability of the twenty most frequent words in the data collected in Malaysia and Singapore.



Corpus Analysis

- Sociolinguistic difference
  - Speakers in Malaysia and Singapore have different preference in selecting words for conversation.
    - For example, Malaysia speakers used “*tu*” (*that*) almost three times more than Singapore speakers.
  - Perplexity evaluation
    - The difference of word distribution in both subsets can be reflected by the perplexity of the unigram language model (as follows).

	Malaysia	Singapore
Malaysia	484.4	780.2
Singapore	1025.8	467.8

Corpus Analysis

- Sociolinguistic difference
  - – The combined corpus is robust to the sociolinguistic variation. The following shows the unigram and bigram perplexities after combining the corpus from both sides.

	Malaysia	Singapore	All
Malaysia	484.4	1025.8	497.2
Singapore	780.2	467.8	502.6
All	565.9	793.9	498.9

	Malaysia	Singapore	All
Malaysia	218.5	754.1	218.8
Singapore	534.6	231.9	234.1
All	293.0	512.3	223.7

Summary

- We have collected and transcribed 50 hours of Malay conversational spontaneous speech. Also additional annotations are tagged to some selected linguistic and acoustic events, in order to facilitate our future research.

- The corpus will be used extensively for our research in conversational LVCSR, in the core area of
  - Pronunciation modeling
  - Disfluencies detection
  - Language modeling
  - Noise compensation

language diarization on code-switch speech

Speaker: dau-cheng lyu

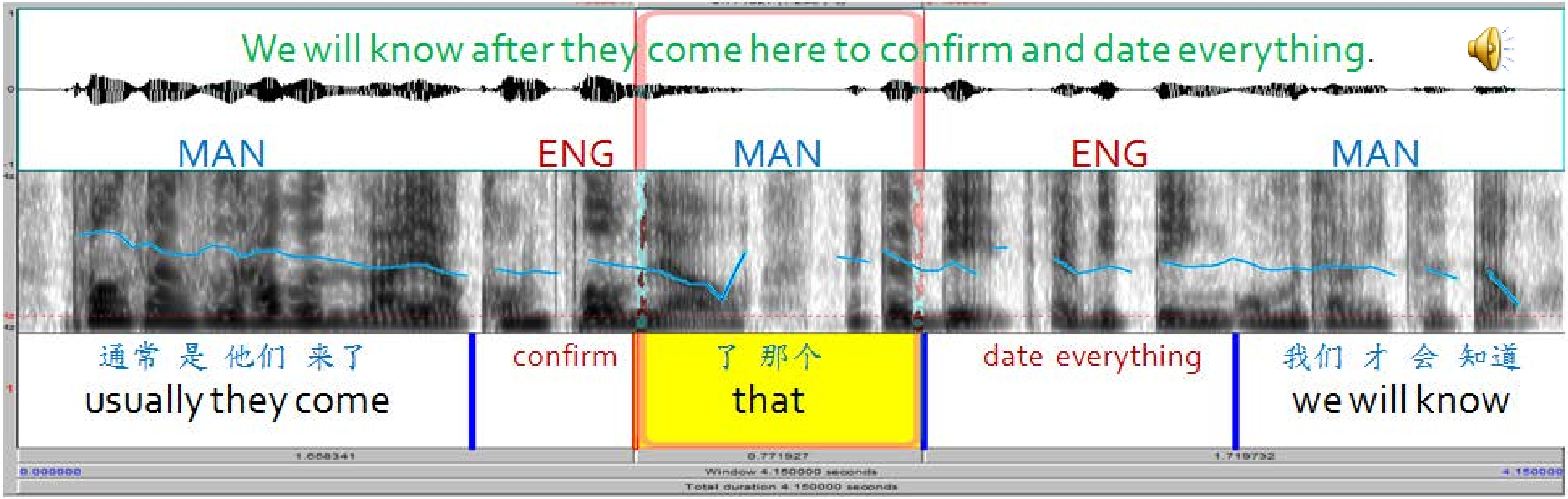
NANYANG
TECHNOLOGICAL
UNIVERSITY

Outline

- **What is code-switch speech**
- **Language recognition V.S. language diarization**
- **Framework for language diarization**
- **Experiment and Result**
- **Conclusion**

What is code-switch speech

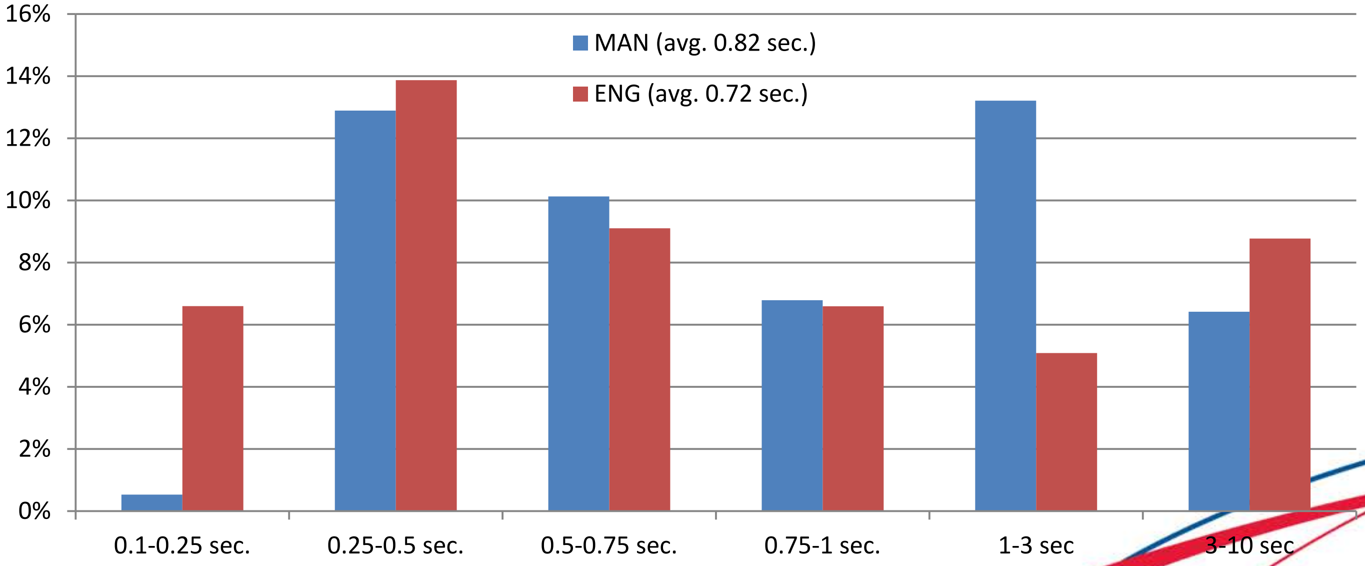
- **Code-switch speech:**
  - **refers to the switching of languages in speech**
    - Mandarin-Taiwanese, Taiwan
    - Mandarin-English, Singapore
    - Cantonese-English, HongKong
    - English-Spanish, U.S.A.



We will know after they come here to confirm and date everything.

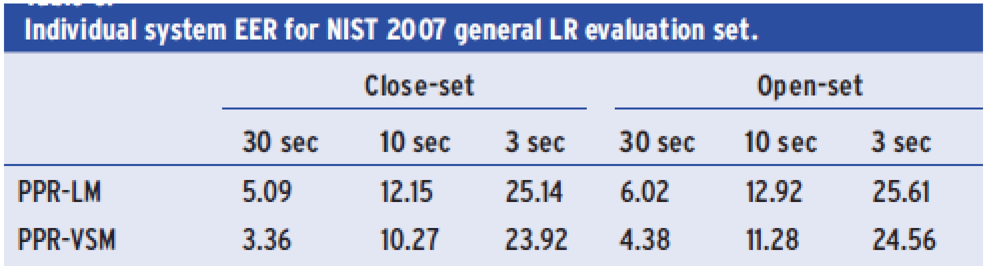
SEAME (South-East Asia English Mandarin) speech corpus

1) 63-hour of conversational code-switch speech
    1) Interview and conversational recording
    2) 16k Hz, 16-bit sampling rate
2) Collected from Singapore and Malaysia
3) Over 80% mono-lingual segments are less one second
4) 2.2 language switching within a utterance



Language Recognition (LID)

- Given an utterance to identify language identity
  - Language identity is unknown
  - But language segmentation is known
- Systems:
  - GMM-UBM, PRLM (Phoneme Recognizer followed by Language Model), P-PRLM (Parallel PRLM), P-PR-SVM and PPR (Parallel Phone Recognition),
- Features:
  - Acoustic, prosodic, phonotactic, lexical and syntactic
- Performance:



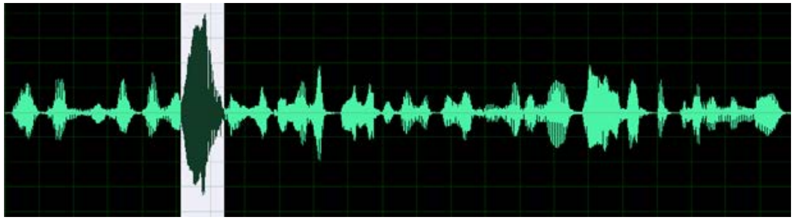
Individual system EER for NIST 2007 general LR evaluation set.

	Close-set			Open-set		
	30 sec	10 sec	3 sec	30 sec	10 sec	3 sec
PPR-LM	5.09	12.15	25.14	6.02	12.92	25.61
PPR-VSM	3.36	10.27	23.92	4.38	11.28	24.56

E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu,
"Language Identification: A Tutorial", in IEEE Circuits and Systems Magazine,
Volume: 11, Issue: 2, pp.82-108, 2011

Language Diarization

- **Task definition**
  - Given a code-switch utterance to estimate language identity and language segmentation
    - Both language identities and segmentation are unknown
  - Similar to “speaker diarization”
    - The task of determining "who spoke when?"
- **Challenges:**
  - the start and end points of a language segment is unknown;
  - language segments are much shorter than those studied in the prior work



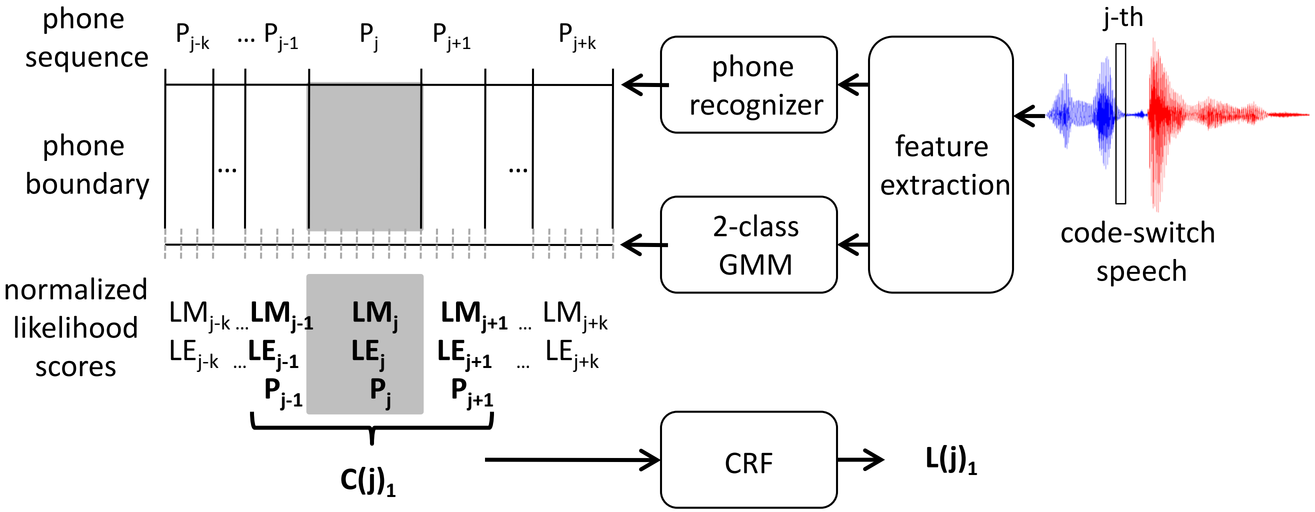
我 讲 是 他 讲 的 [啦] then 他 就 not happy 他 就 讲 我 put words in 他 的 mouth then after that amanda

Language diarization on code-switch utterance

- **One of the goals**
  - provide a language likelihood score as a soft language decision to support multilingual speech recognition system
- **Analysis on code-switch utterances**
  - Very short monolingual segment
  - code-switch often occur between word, phrase or sentence
- **Idea**
  - Combining acoustic and phonotactic feature
    - Complementary features for language recognition
  - Using segmental based language recognition
  - Using context information for back-end classifier

NANYANG TECHNOLOGICAL UNIVERSITY

language diarization system for code-switch utterance (fusion system)


phone sequence
phone boundary
normalized likelihood scores
P_{j-k} ... P_{j-1} P_j P_{j+1} P_{j+k}
...
...
LM_{j-k} ... LM_{j-1} LM_j LM_{j+1} ... LM_{j+k}
LE_{j-k} ... LE_{j-1} LE_j LE_{j+1} ... LE_{j+k}
P_{j-1} P_j P_{j+1}
$C(j)_1$
phone recognizer
2-class GMM
feature extraction
j-th
code-switch speech
CRF
$L(j)_1$

Front-end classifier: GMM and phone recognition
Back-en classifier: CRF (conditional random fields)

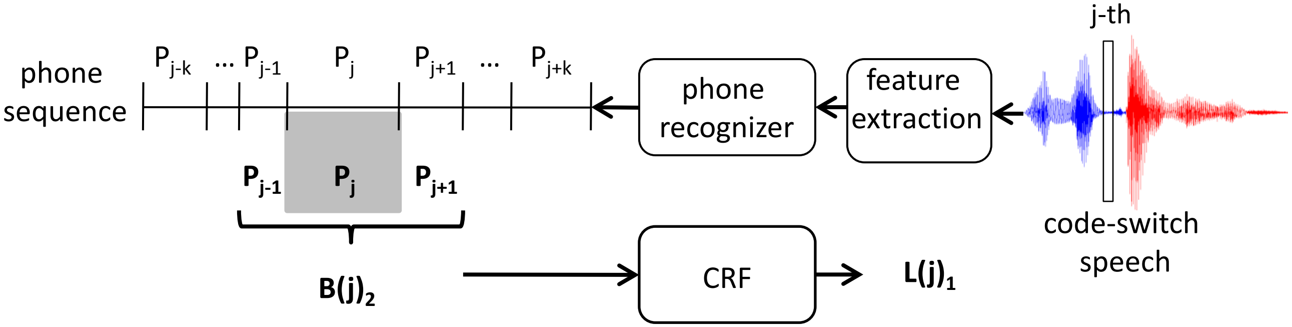
Features for front end: LDA42
Features for back-end: phone-based segment acoustic and phonetic features (LMj, Lej and Pj)

Temporal features: concatenating variant length of left and fight combined feature,
C(j)k, j: the j-th phone segment,
k: the number of segments used in the left and right context

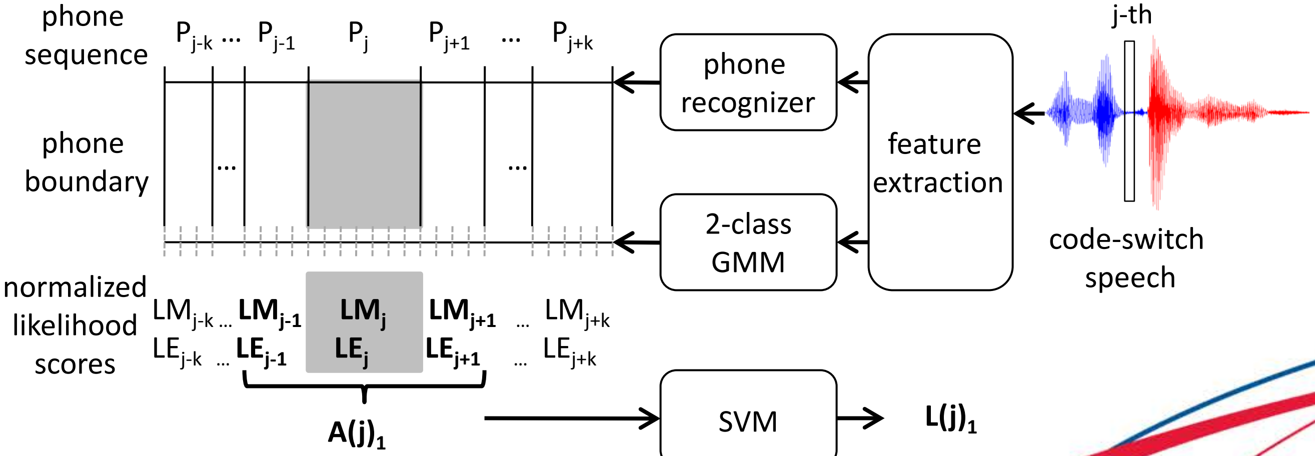
NANYANG TECHNOLOGICAL UNIVERSITY

language diarization system for code-switch utterance (individual system)

- **Phonotactic language diarization system**



- **Acoustic language diarization system**


j-th
phone sequence
P_{j-k} … P_{j-1} P_j P_{j+1} … P_{j+k}
phone boundary
…
…
phone recognizer
feature extraction
2-class GMM
code-switch speech
normalized likelihood scores
LM_{j-k} … LM_{j-1} LM_j LM_{j+1} … LM_{j+k}
LE_{j-k} … LE_{j-1} LE_j LE_{j+1} … LE_{j+k}
$A(j)_1$
SVM
$L(j)_1$

NANYANG TECHNOLOGICAL UNIVERSITY

Corpus and experiential settings

- **SEAME corpus**

	# of speakers	# of utterances	# of hours	MAN(%)	ENG(%)	SIL(%)	Others(%)
training set	133	44,524	52.38	45	27	20	8
dev. set	11	3,505	5.26	45	25	20	10
test set	13	4,116	5.21	44	27	20	9

•**Feature**: LDA42, contains 5 continuous frames MFCC-based feature reduced dimension by using linear discriminant analysis (LDA) to 42

•**Training data**: used to train front-end classifier, e.g. GMM and phone recognizer

•**Dev. data**: used to train back-end classifier, e.g. SVM, CRF

•**VAD** (voice activity detection) is used to identify speech and silence segment. Performance of current VAD is 5.88% frame error rate.

LID classifier: two classes (Mandarin and English)

Others: will not be correctly identified (Other Languages, Discourse Particles, Other sounds, Fillers)

NANYANG TECHNOLOGICAL UNIVERSITY

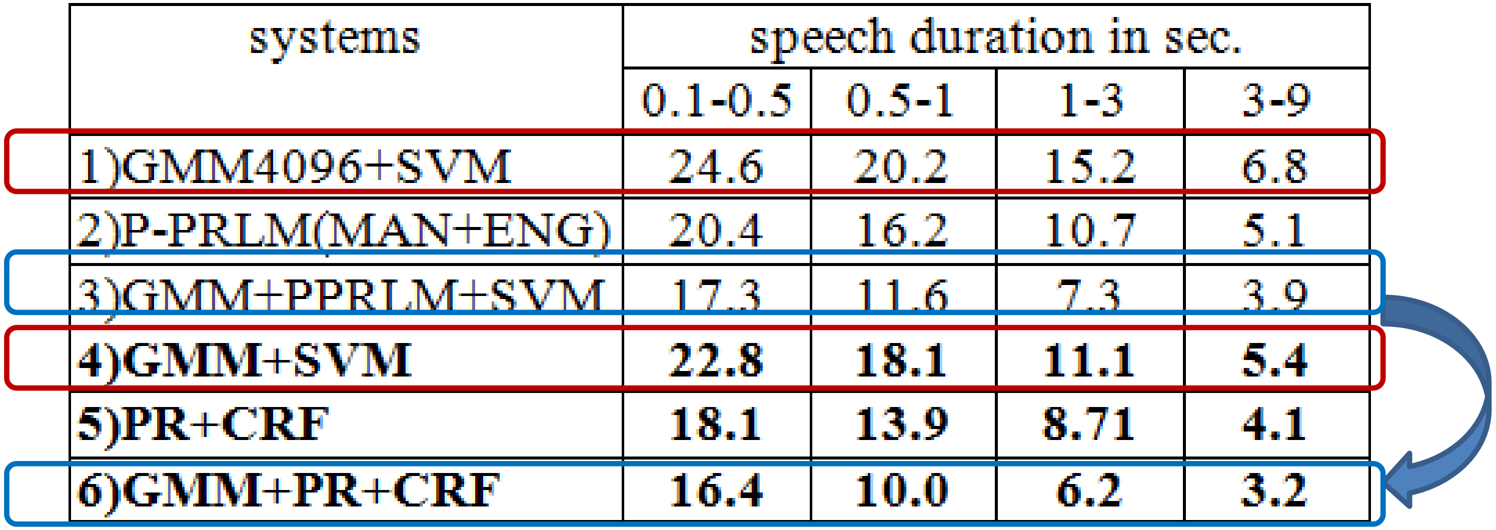
Two tasks

- **Language recognition**
  - Test utterance:
    - Monolingual speech segment manually extracted from SEAME corpus
  - Evaluation: equal error rate (EER)
  - Comparison:
    - GMM, P-PRLM, GMM+PPRLM+SVM
- **Language diarization**
  - Test utterance:
    - Code-switch speech
  - Evaluation: frame error rate (FER)
    - convert the current language identity for each phone segment into frame level

FER: frame error rate (%)

Performance for language recognition

- **Divided test data into for categories according to duration**
- **Obtained relative equal error rate reduction of 5.2%, 13.8%, 15.1% and 17.9%**



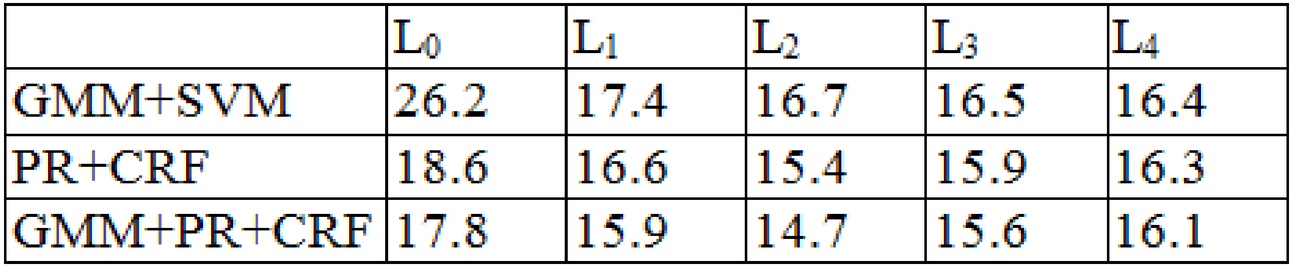
systems	speech duration in sec.			
	0.1-0.5	0.5-1	1-3	3-9
1)GMM4096+SVM	24.6	20.2	15.2	6.8
2)P-PRLM(MAN+ENG)	20.4	16.2	10.7	5.1
3)GMM+PPRLM+SVM	17.3	11.6	7.3	3.9
4)GMM+SVM	**22.8**	**18.1**	**11.1**	**5.4**
5)PR+CRF	**18.1**	**13.9**	**8.71**	**4.1**
6)GMM+PR+CRF	**16.4**	**10.0**	**6.2**	**3.2**

Other state of the art method

Proposed framework

Performance for language diarization

- To evaluate the effect of segment duration to performance, we use variant length of context information and the corresponding LID outputs are L_0, L_1, L_2, L_3, and L_4
- Longer temporal information of features for back-end perform better
- Combined feature (acoustic + phonotactic) performs best



	L_0	L_1	L_2	L_3	L_4
GMM+SVM	26.2	17.4	16.7	16.5	16.4
PR+CRF	18.6	16.6	15.4	15.9	16.3
GMM+PR+CRF	17.8	15.9	14.7	15.6	16.1

Conclusion

- **introduce the language diarization task on code-switch utterances**
- **Performance improvement**
  - Proposed temporal features for back-end classifier to process very short monolingual segment in code-switch speech
  - Combined acoustic and phonotactic features
- **Proposed framework outperforms other state-of-the-art language recognition system**
- **Achieved 14.7% frame error rate on language diarization task**

AN ANALYSIS OF VECTOR TAYLOR SERIES MODEL COMPENSATION FOR NON-STATIONARY NOISE IN SPEECH RECOGNITION
(Noise Conditioning – VTS method)

presented by

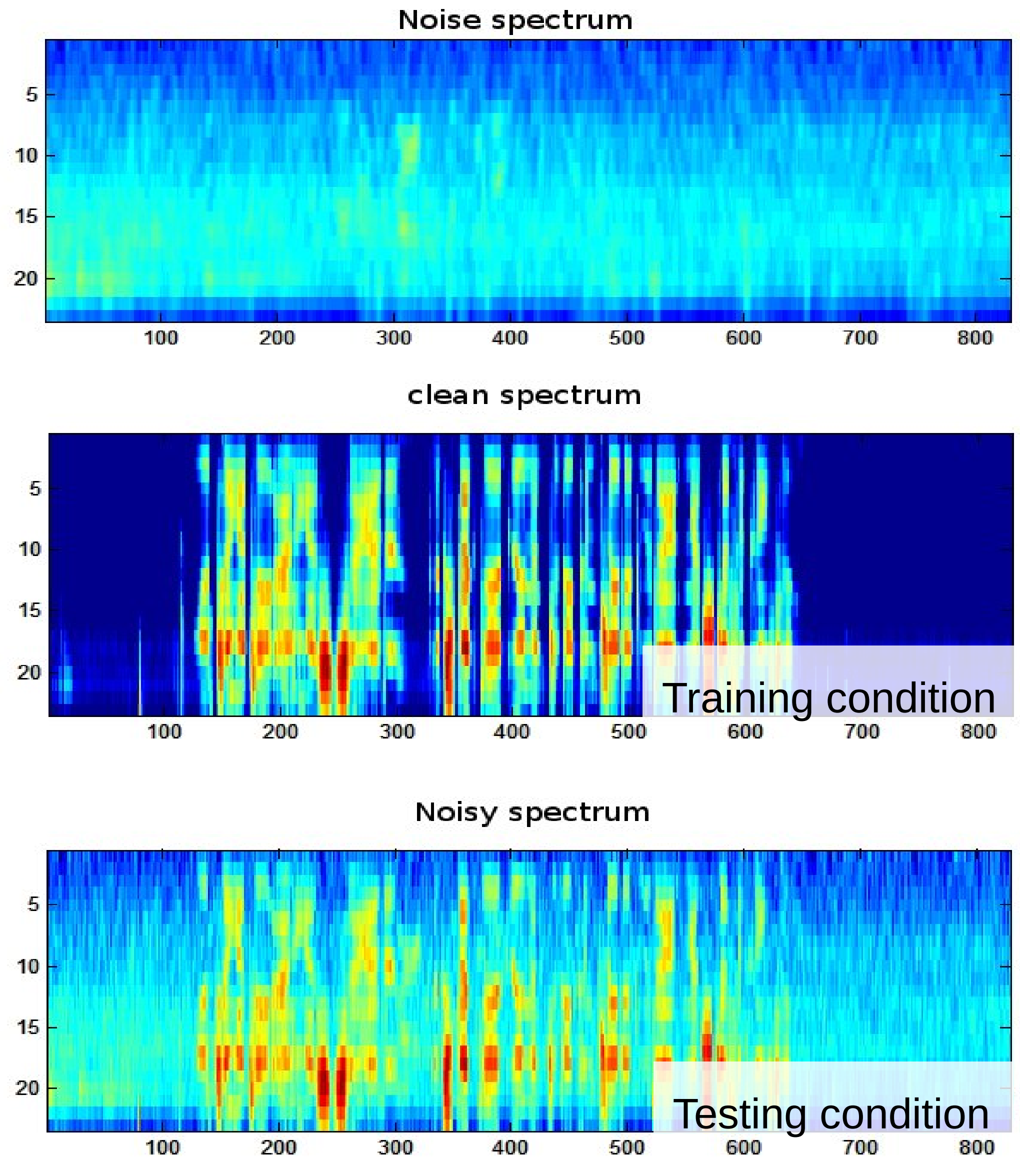
Duc Hoang Ha Nguyen
School of Computer Engineering,
Nanyang Technological University, Singapore

23 Nov 2012

Outline

- Background
  - Noise robust speech recognition
- Motivation
  - Noise conditioning approach
- Noise conditioning (NC) – vector Taylor series (VTS) method
- Experiments
  - Aurora2 task
- Summary

Noise robust speech recognition



Problem: mismatch between training and testing conditions

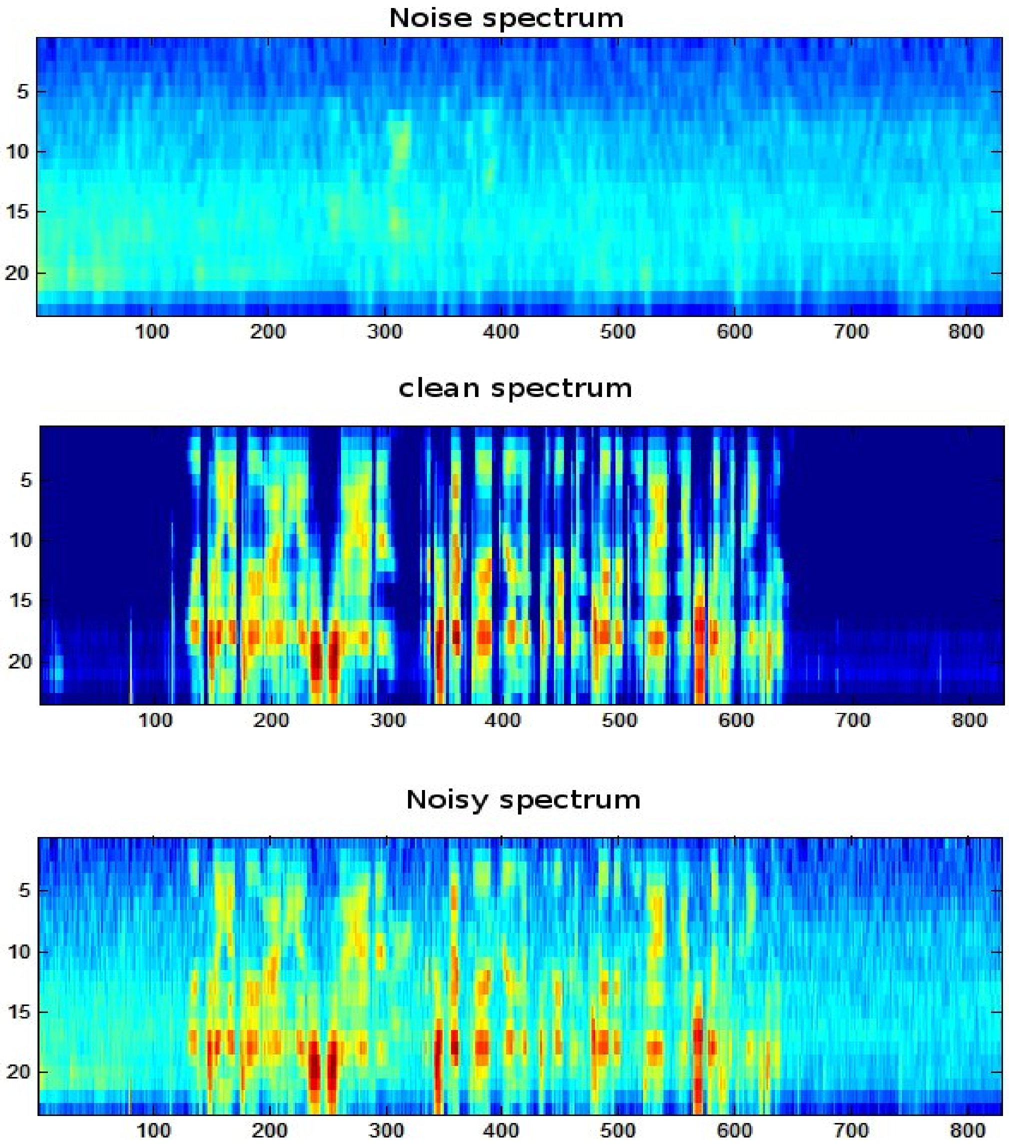
Feature-based approach:
(modify feature to reduce mismatch)
+ Feature normalization:
CMN (Atal 1974)
CVN (Molau 2003)
+ Clean speech estimation
Spectral subtraction (Boll 1979)
MMSE (Ephraim and Malah, 1984)

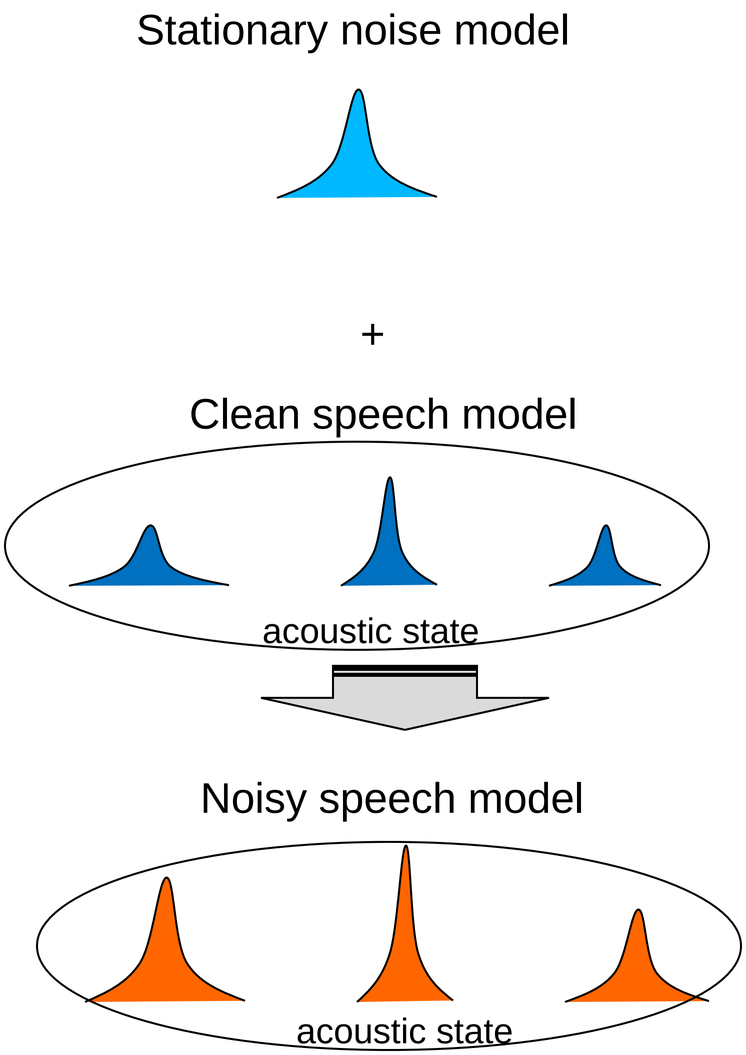
Model-based approach:
(estimate noisy acoustic model
to represent better noisy features)
+ Adaptive approach:
MAP (Gauvain and Lee 1994)
MLLR (Leggetter and Woodland 1995)
+Predictive approach:
PMC (Gales 1995)
VTS (Acero et al. 2000, Jinyu Li 2009)

Motivation

- In this presentation, I will present an analysis of vector Taylor series (VTS) model compensation for non-stationary noise in speech recognition.
- Reason: VTS method can handle well stationary noise, but suffer in non-stationary noisy environment.

Motivation – VTS model compensation approach





Stationary noise model

+

Clean speech model

acoustic state

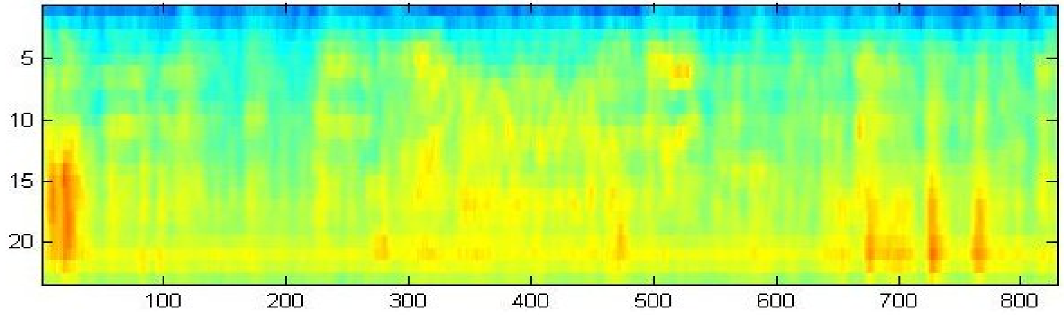
Noisy speech model

acoustic state

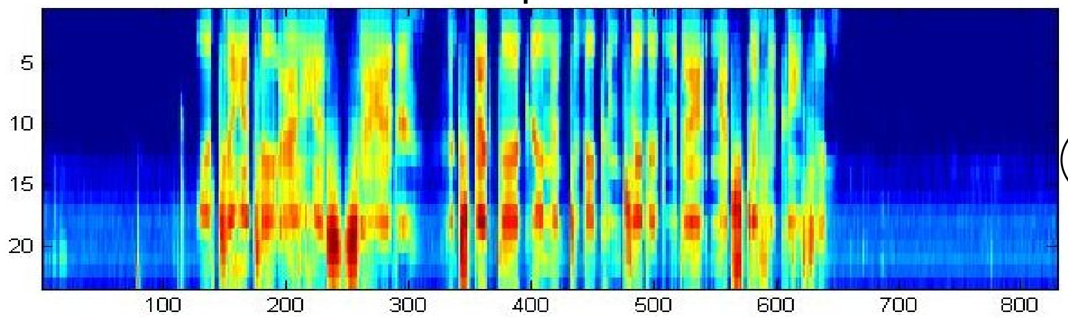
(Moreno 1996, Gales 1998, Acero et al. 2000, Jinyu Li et al. 2009)

Motivation – Issue in non-stationary noisy environments

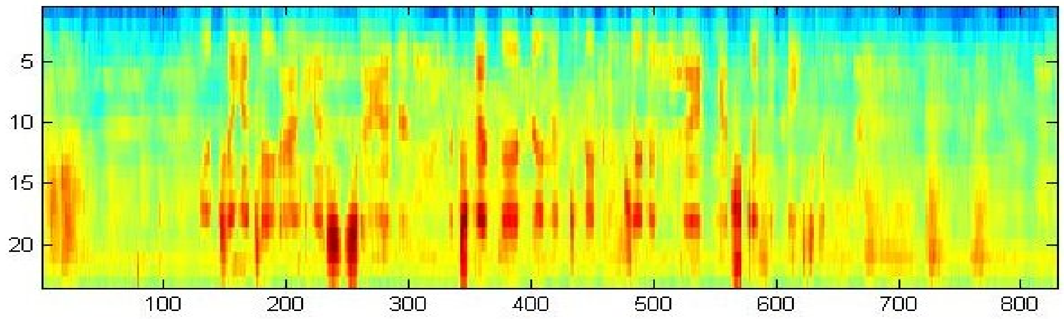
Noise spectrum

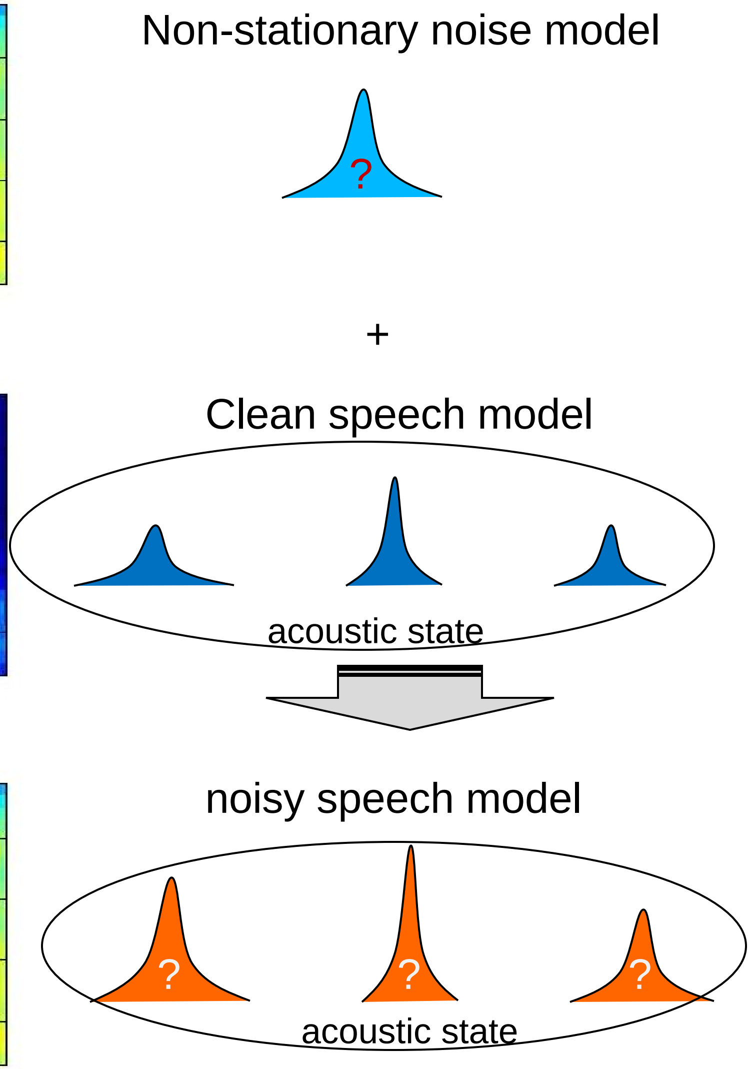


Clean spectrum



Noisy spectrum





Non-stationary noise model

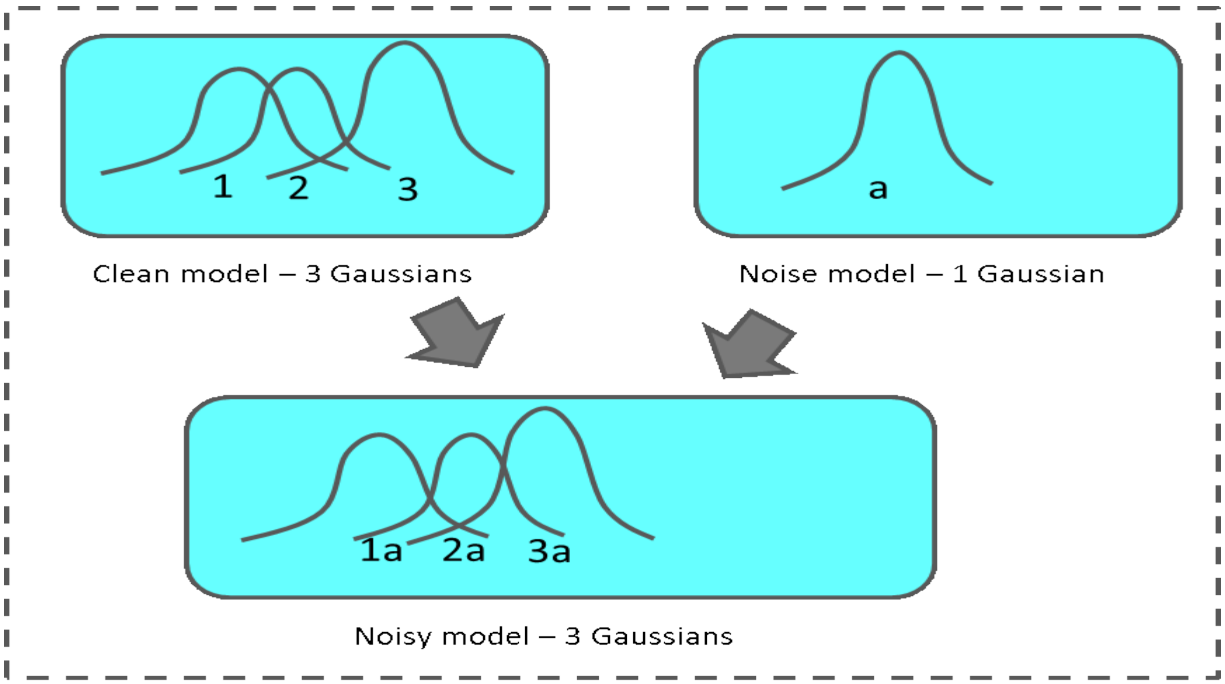
+

Clean speech model

noisy speech model

How to choose noise model?

Noise Gaussian Mixture Model



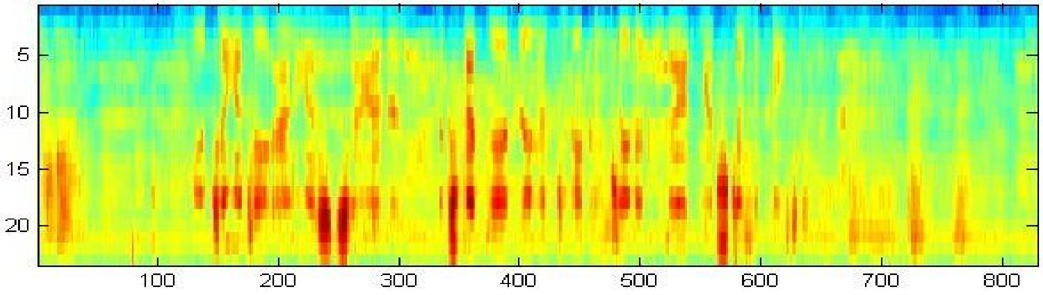
Model combination

(Gales 1998)

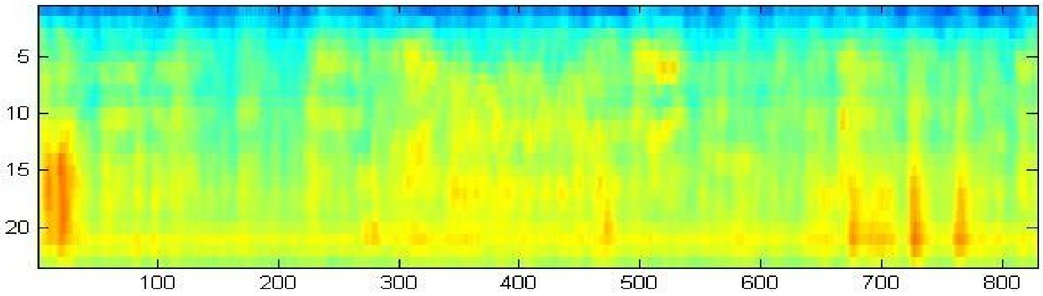
- Complexity significantly increasing
- Extra memory for acoustic model
- Extra computation in decoding
- Hard to optimize noise GMM parameters
- Confusing and not optimal to use all noise mixtures, e.g. babble noise and car noise → babble noise mixture may confuse the ASR system where only have car noise.

Proposed method - Noise conditioning approach

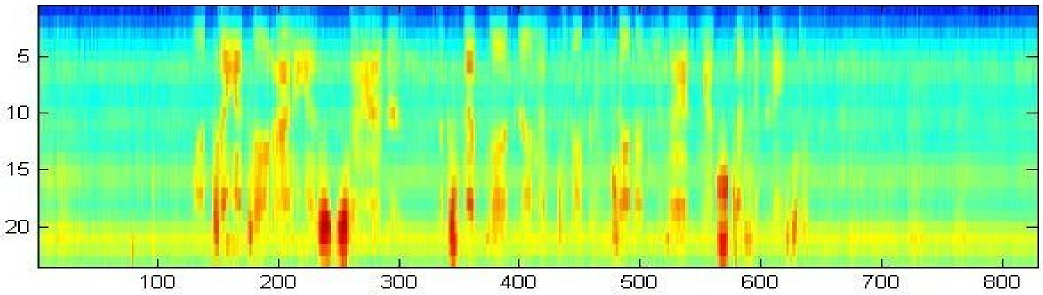
Noisy spectrum



Noise spectrum



Ideal enhanced noisy spectrum



$\boldsymbol{y}_t^{\mathrm{FB}}$: noisy mel filter bank (FB)

noise   Noise mean   non-stationary component

$$\boldsymbol{n}_t^{\mathrm{FB}} = \boldsymbol{\mu}_n^{\mathrm{FB}} + \boldsymbol{z}_t + \boldsymbol{\epsilon}_t$$

Residual error

Enhanced noisy mel FB   Confidence score of noise estimate   Estimated non-stationary component

$$\hat{\boldsymbol{y}}_t^{\mathrm{FB}} = |\boldsymbol{y}_t^{\mathrm{FB}} - \alpha \hat{\boldsymbol{z}}_t|$$

(Nguyen et al. 2012)

Proposed method - Noise conditioning approach

Noise conditioning approach

+ Try to reduce non-stationary characteristic of noise,
i.e. ideal residual noise is stationary.

+ Residual noise is effectively handled by VTS method.

+ If noise estimate is wrong (or confidence score is low),
should not enhance the features,
i.e. keep the observed features.

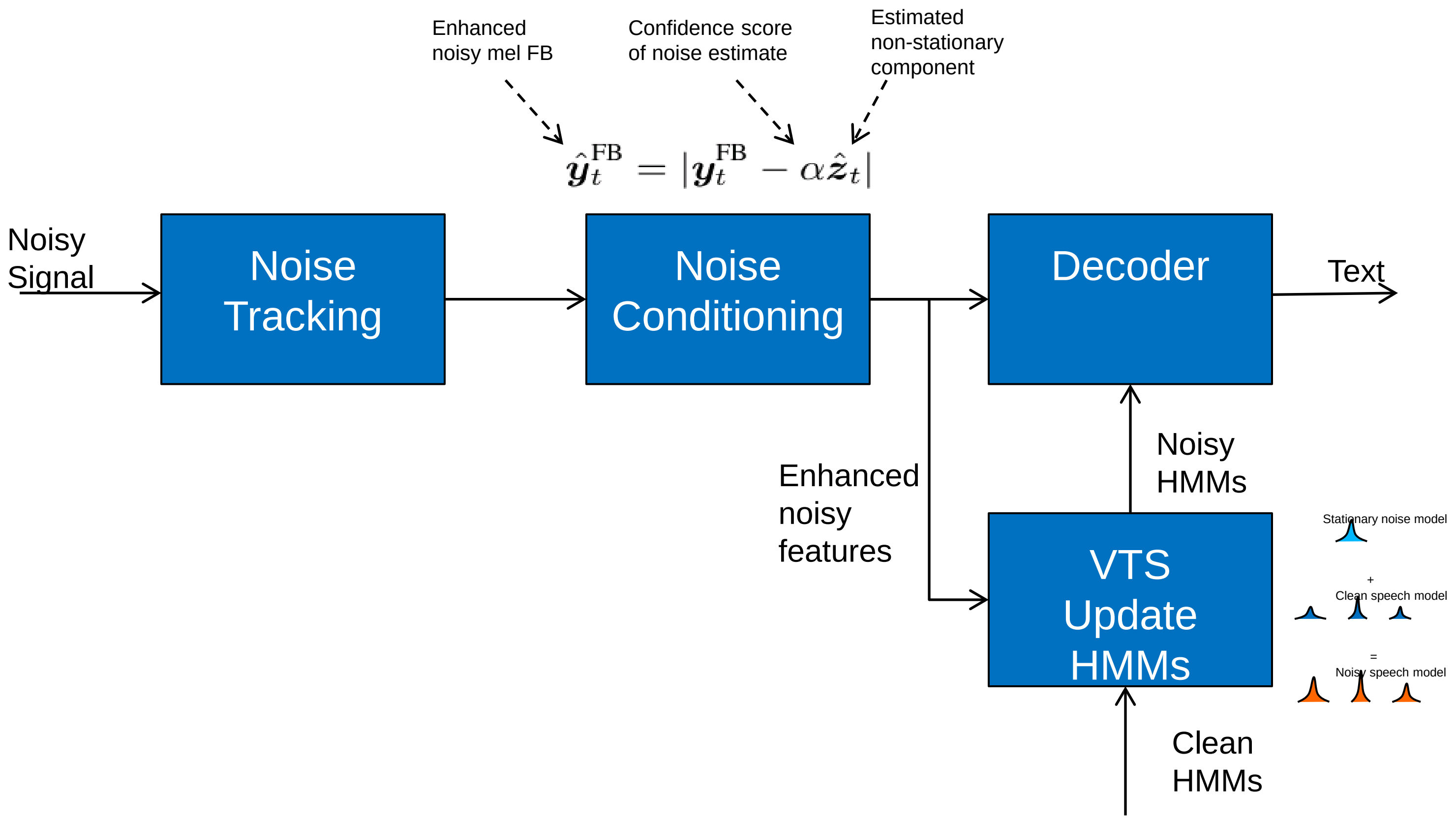
Vs.

Clean features estimation approach

+ Try to estimate clean feature,
i.e. ideal residual noise is null.

- If noise estimate is wrong,
we *do not know what to do*,
i.e. it is hard to estimate clean features.

Noise conditioning – VTS framework



Enhanced noisy mel FB

Confidence score of noise estimate

Estimated non-stationary component

$$\hat{\boldsymbol{y}}_t^{\mathrm{FB}} = |\boldsymbol{y}_t^{\mathrm{FB}} - \alpha \hat{\boldsymbol{z}}_t|$$

Noisy Signal

Noise Tracking

Noise Conditioning

Decoder

Text

Enhanced noisy features

Noisy HMMs

VTS Update HMMs

Clean HMMs

Stationary noise model

+

Clean speech model

=

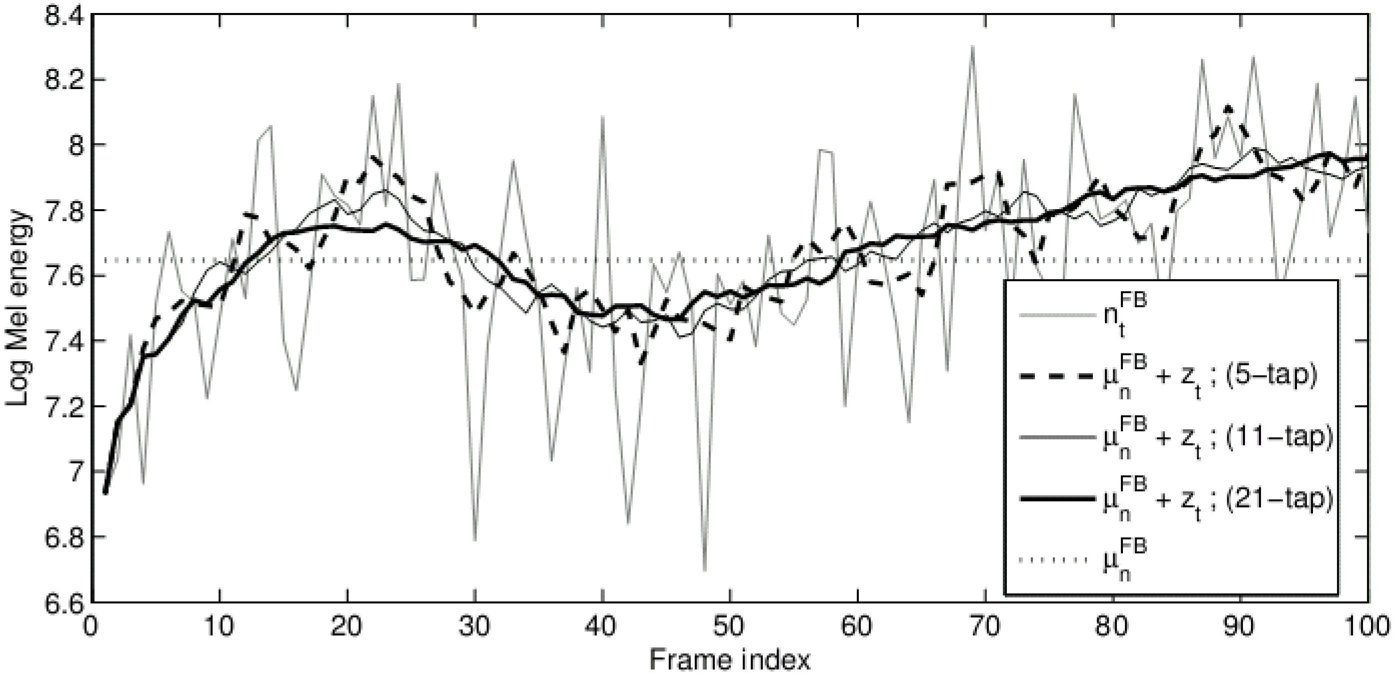
Noisy speech model

Experiments

- Noise estimate simulation

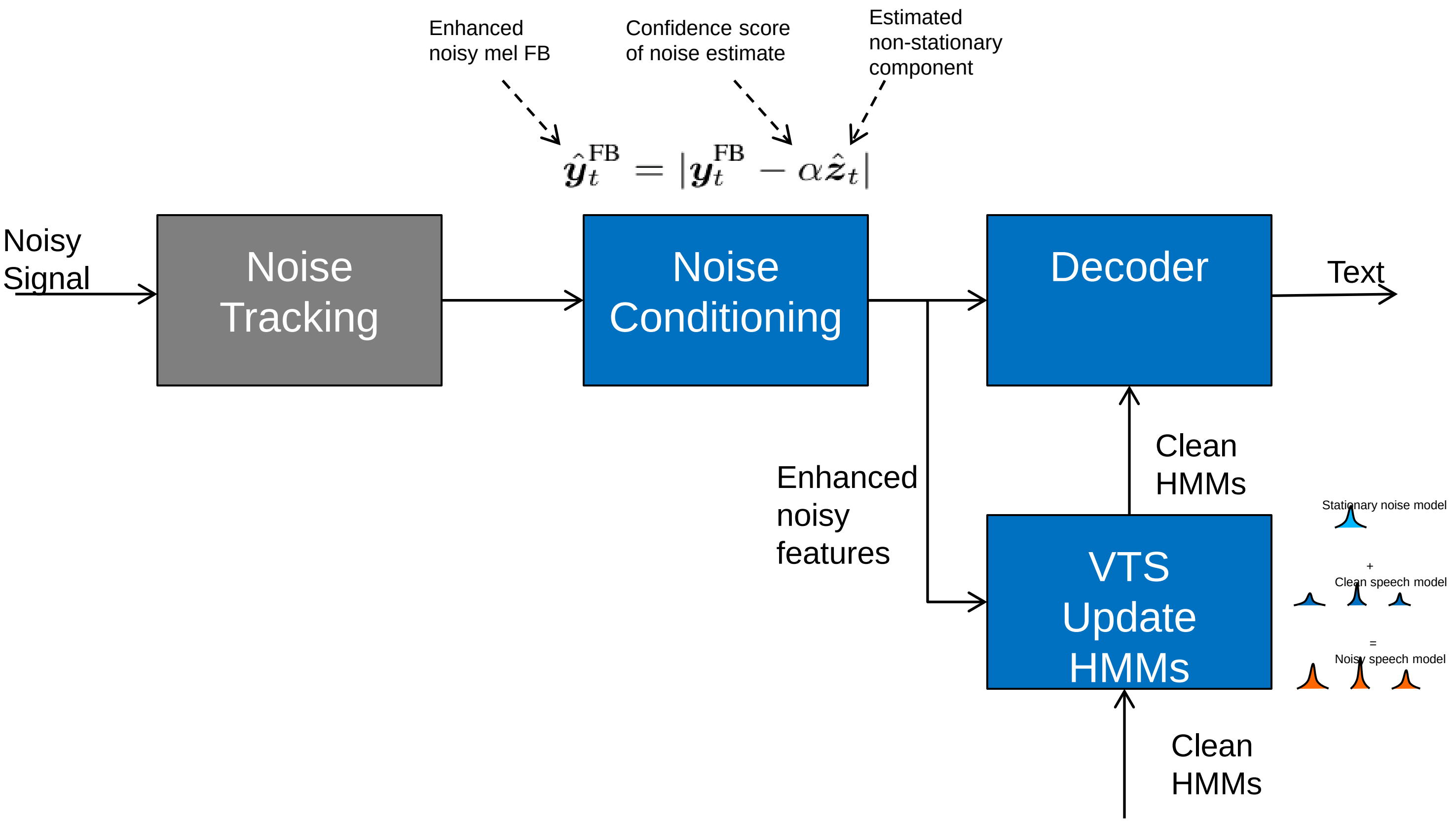
  (To isolate the effects of noise estimate in this examination)
- Noise model for VTS method with/without the noise conditioning process
- Speech recognition performance on AURORA 2 task
- Aurora 2 task: English spoken digit data

Experiment - Noise estimate simulation



- Noise estimate is a smoothed version of true noise features.
- Noise estimate captures the trend of the noise features.
- The purpose is to isolate the effects of noise estimate in this examination.

Noise conditioning – VTS framework

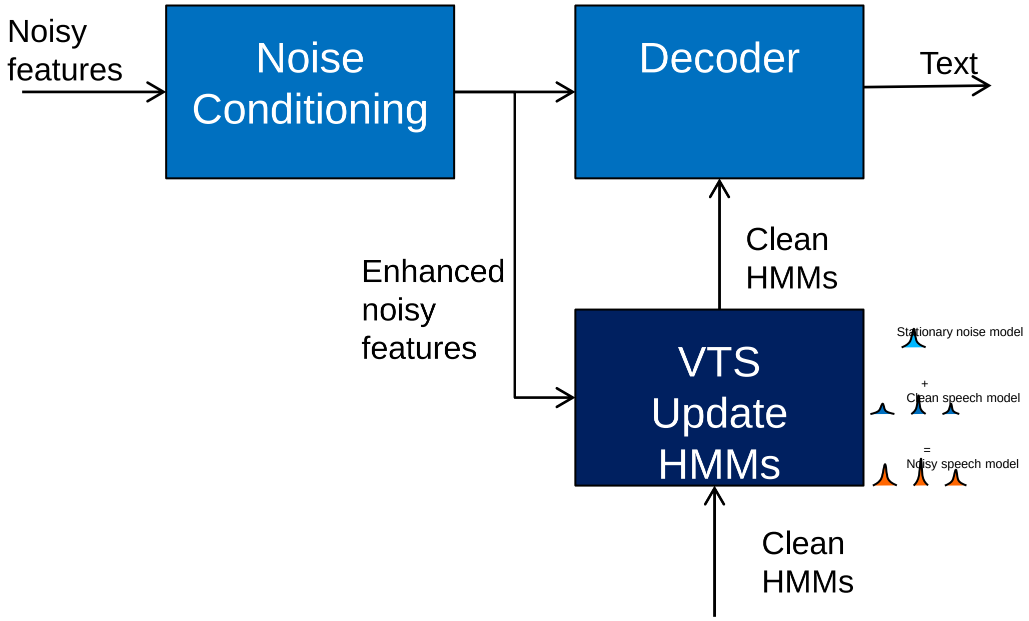


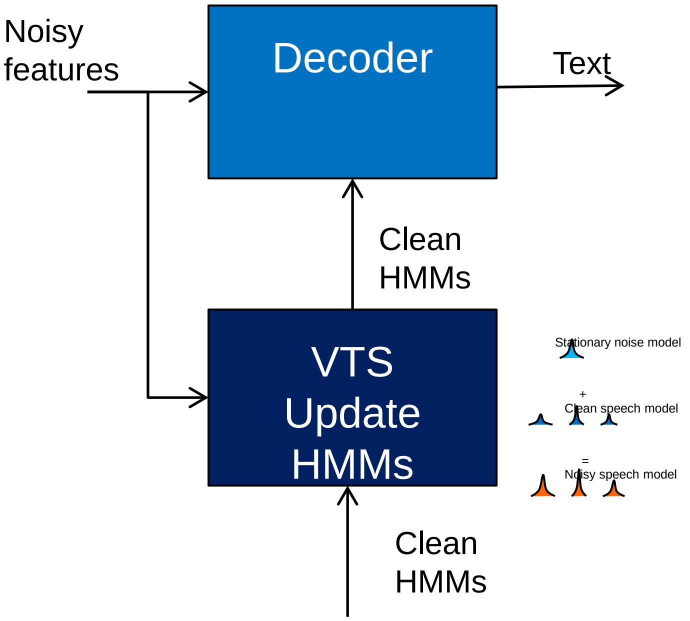
$$\hat{\boldsymbol{y}}_t^{\mathrm{FB}} = |\boldsymbol{y}_t^{\mathrm{FB}} - \alpha \hat{\boldsymbol{z}}_t|$$

(Focus on analysis of effects of noise conditioning)

Baseline Comparison

NC-VTS Vs. VTS





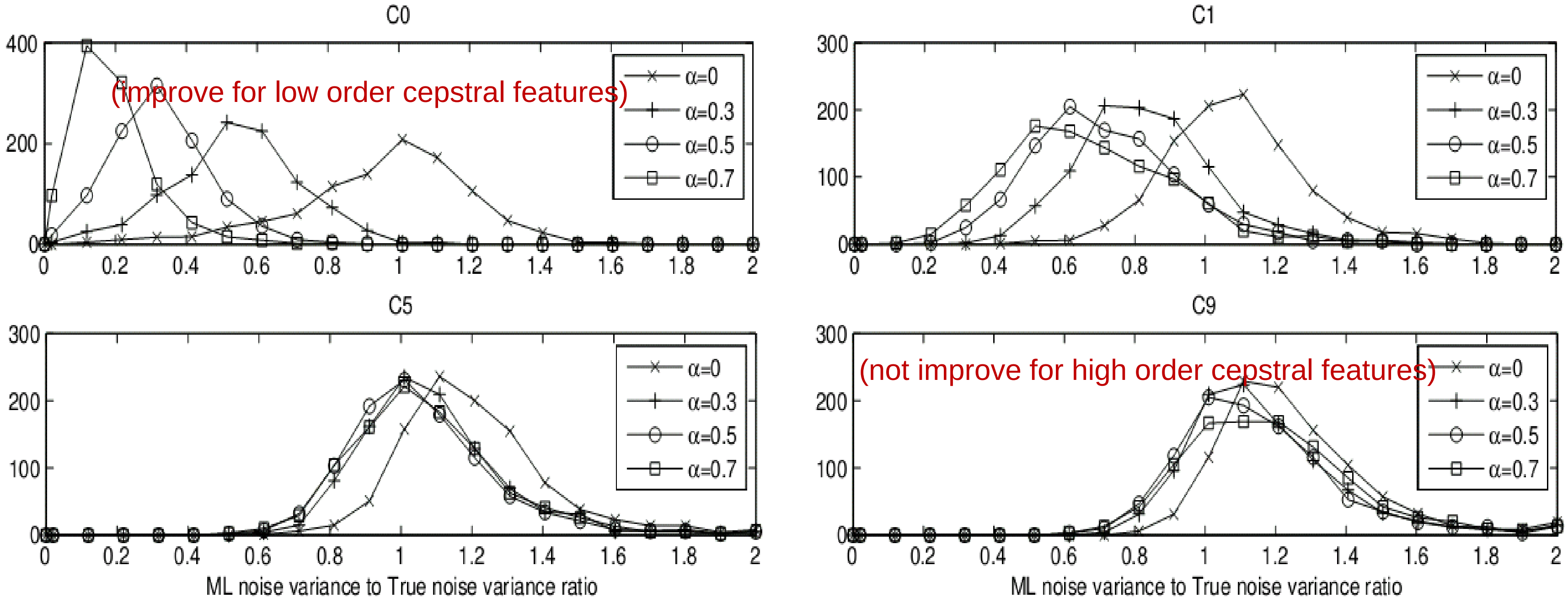
Considering model adaptation, the major difference is the noise model.
Noise model is estimated using maximum likelihood algorithm [Jinyu Li et al., 2009]
The more stationary the noise, the smaller the variance is.

Experiments – Non-stationary characteristic Analysis

Enhanced noisy mel FB

Estimated non-stationary component

$$\hat{\boldsymbol{y}}_t^{\text{FB}} = |\boldsymbol{y}_t^{\text{FB}} - \alpha \hat{\boldsymbol{z}}_t|$$



(improve for low order cepstral features)

(not improve for high order cepstral features)

Histogram of ratio (Estimated noise variance / True noise variance)
Noise model is estimated using maximum likelihood algorithm [Jinyu Li et al., 2009]
(Smaller is better)

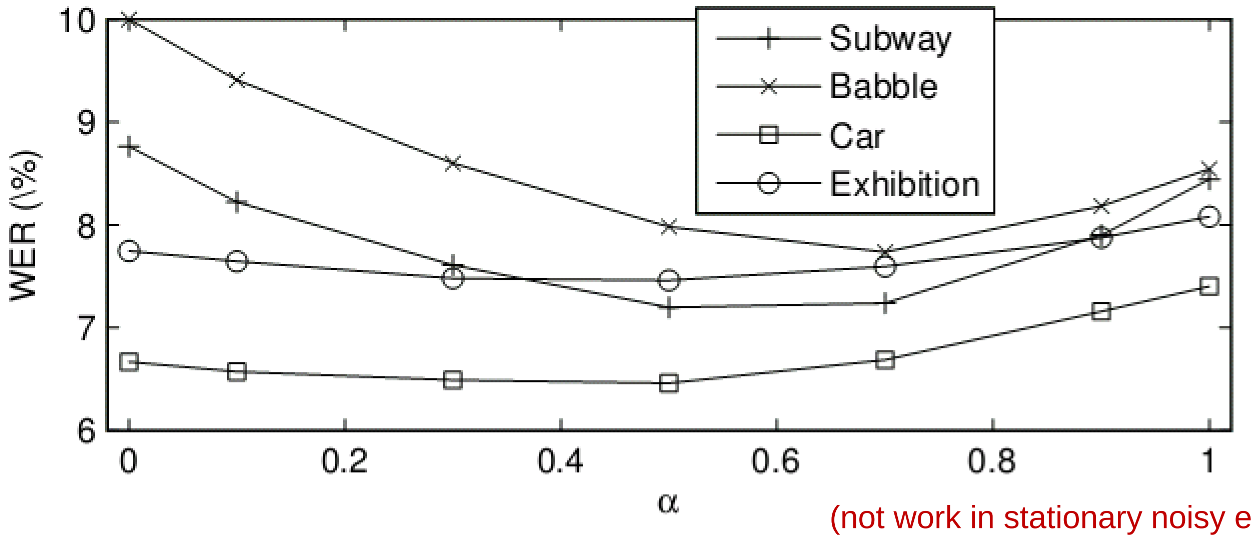
Experiments – Speech Recognition Performance

Enhanced noisy mel FB

Estimated non-stationary component

$$\hat{\boldsymbol{y}}_t^{\mathrm{FB}} = \left|\boldsymbol{y}_t^{\mathrm{FB}} - \alpha \hat{\boldsymbol{z}}_t\right|$$

(significantly improve performance in non-stationary noisy environment)



(not work in stationary noisy environment)

AURORA 2 task
Training data: clean set
Testing data: test set A
Noise estimate simulation: smooth win-size=11 frames

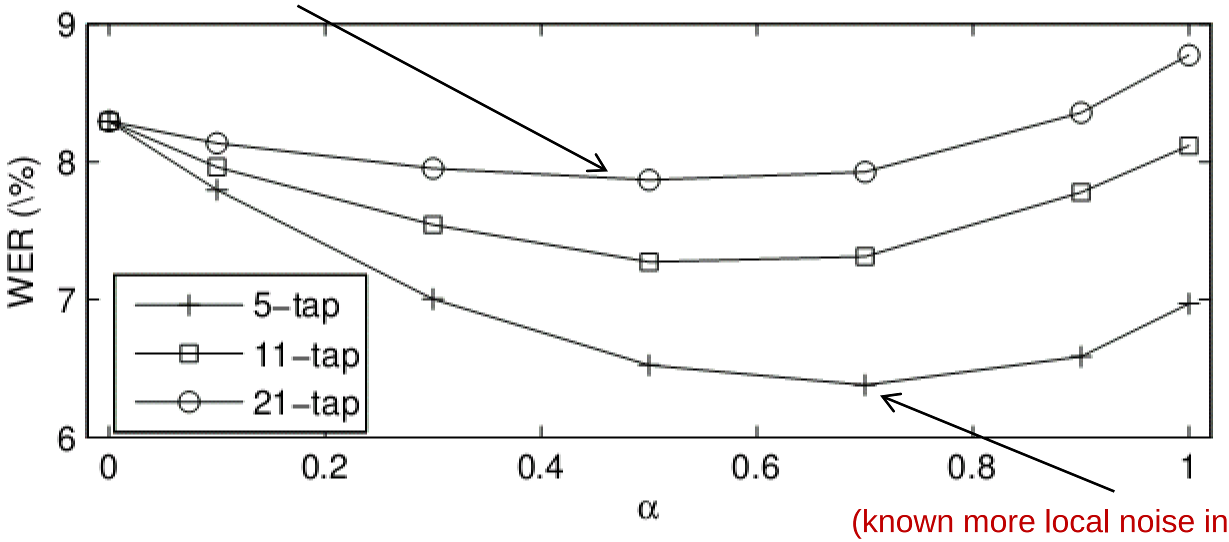
Experiments – Speech Recognition Performance

Enhanced noisy mel FB

Estimated non-stationary component

$$\hat{\boldsymbol{y}}_t^{\mathrm{FB}} = \left|\boldsymbol{y}_t^{\mathrm{FB}} - \alpha \hat{\boldsymbol{z}}_t\right|$$

(known only the trend of the noise features)



(known more local noise information)

AURORA 2 task
Training data: clean
Testing data: test set A
Noise estimate simulation: smooth win-size=5,11,21 frames

Summary

- Present an analysis of noise conditioning method that aims to reduce the non-stationary characteristics of the noise,
- Use noise conditioning method as a preprocessor of the VTS model compensation method,
- Investigate the noise robustness potential of the noise conditioning –VTS method in non-stationary noisy environments.
- Future work will focus on non-stationary noise tracking

References

P. J. Moreno, Speech Recognition in Noisy Environments. PhD thesis, Carnegie Mellon University, 1996.

M. J. F. Gales, “Predictive model-based compensation schemes for robust speech recognition,” Speech Communication, vol. 25, pp. 55–64, 1998.

A. Acero, L. Deng, T. Kristjansson, and J. Zhang, “HMM adaptation using vector Taylor series for noisy speech recognition,” in Proc. Intl. Conf. on Spoken Language Processing, 2000, pp. 869–872.

J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, “A unified frame-work of HMM adaptation with joint compensation of additive and convolutive distortions,” Computer Speech and Language, vol. 23, pp. 389–405, July 2009.

D. H. H. Nguyen, X. Xiao, E. S. Chng, H. Li, “*An analysis of vector Taylor series model compensation for non-stationary noise in speech recognition*”, in Proc. ISCSLP, 2012.

Thank you!