INTERSPEECH 2009 BRIGHTON

# Dynamic Features in the Linear Domain for Robust Automatic Speech Recognition in a Reverberant Environment

*Osamu Ichikawa, Takashi Fukuda, Ryuki Tachibana, Masafumi Nishimura*

Tokyo Research Laboratory, IBM Research

{ichikaw, fukuda1, ryuki, nisimura}@jp.ibm.com

## Abstract

Since the MFCC are calculated from logarithmic spectra, the delta and delta-delta are considered as difference operations in a logarithmic domain. In a reverberant environment, speech signals have trailing reverberations, whose power is plotted as a long-term exponential decay. This means the logarithmic delta value tends to remain large for a long time. This paper proposes a delta feature calculated in the linear domain, due to the rapid decay in reverberant environments. In an experiment using an evaluation framework (CENSREC-4), significant improvements were found in reverberant situations by simply replacing the MFCC dynamic features with the proposed dynamic features.

**Index Terms**: automatic speech recognition, dynamic feature, reverberation, linear delta, delta, MFCC

10.21437/Interspeech.2009-9

## 1. Introduction

Many approaches have been made to improve the accuracy of automatic speech recognition (ASR) in reverberant environments. They can be classified into these categories:

(A) Front end processing
(B) Multi-condition training
(C) Adaptation
(D) Features

Category (A) compensates for reverberation by preprocessing the speech input. This includes power-envelop restoration [1], harmonic-structure-based filtering [2], linear predictive filtering [3], microphone array technologies [4]-[6], etc. Category (B) trains the acoustic models with reverberant speech data [7]-[9]. Techniques in (C) are adaptations to convert observed features or acoustic models to match each other. The fMLLR [10] is a feature conversion technique, while HMM composition [11] and MLLR [12] are model conversion techniques. Category (D) covers new features that are robust against reverberation. This includes LDA derived RASTA features [13] and the Kernel PCA [14]. For additional improvements, most of these techniques can be combined across categories, and so technology advances in all categories are desired.

Our proposal is in Category (D). We describe the design of new dynamic features that are robust in reverberant environments. The static part can be any existing features such as mel-frequency cepstral coefficients (MFCC).

In many of the current ASR systems, the dynamic features (delta and delta-delta) of MFCC [15] are still used along with the static features of the MFCC. Since the MFCC is calculated from logarithmic spectra, its delta and delta-delta are considered as logarithmic difference operations.

With this background, Figure 1 shows our motivation. If a current phoneme has any frequency region (such as Ω) where spectral power is sparse, that region is susceptible to a preceding phoneme's reverberation. Since reverberation tends to decay exponentially, the spectral power in the region can presumably be plotted as a long constant slope downwards on a logarithmic scale. This means the delta feature keeps the trailing value for a long time, as long as the deltas are logarithmic. That leads to higher error rates in the ASR. In contrast, the gradient of the slope quickly diminishes if the power is plotted on a linear scale. This is our motivation to use the delta values in a linear domain so as to minimize the affects of the preceding phoneme's reverberation.

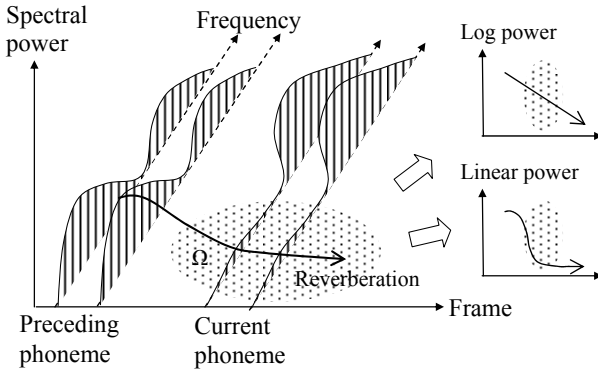


Figure 1: Gradient of spectral power in reverberant region, plotted both in logarithmic scale and in linear scale.

## 2. Dynamic features in a linear domain

It is a simple idea to calculate the differences of spectral power or magnitude in linear domains instead of logarithmic domains. However, this approach has not been widely adopted, because naive implementations raise several concerns.

One concern is that the human auditory system has a logarithmic response following the Weber-Fechner law, so using raw linear deltas may not conform to this law. An alternative approach would use the logarithms after finding the deltas in a linear domain. However, this raises issues of how to handle any negative values as logarithms.

The second concern is that values in a linear domain have wider dynamic ranges due to input gain variations. Delta values in a linear domain need some normalization for statistical modeling.

When we take differences in a logarithmic domain, the multiplicative parts can be cancelled automatically. This works in the same way as cepstral mean subtraction (CMS), and it has a strong advantage over the linear domain operation. Therefore, the third concern is that we may lose chances to compensate for channel distortion.

To address some of these concerns, we propose an approach to divide the spectral difference in a linear domain by the averaged spectrum of the whole utterance. This operation compresses the dynamic range and also it works as channel normalization.

## 2.1. Linear delta feature

Figure 2 shows the base process to generate the linear delta feature proposed.

Firstly, delta spectrum is calculated in a linear domain as Equation (1):

$$\Delta s_t = \frac{\sum_{\theta=1}^{\Theta} \theta\left(s_{t+\theta} - s_{t-\theta}\right)}{2\sum_{\theta=1}^{\Theta} \theta^2}, \quad (1)$$

where $s_t$ is magnitude of the observed spectrum in the t-th frame. It is non-mel and non-logarithmic. The next step is Mel-Filter Bank (Mel-FB) analysis. It accumulates the delta spectrum at each bin with associated triangular weight $W$ in the bank. This outputs the values in the bank as a mel-delta spectrum $\Delta S_t$ in Equation (2):

$$\Delta S_t(l) = \sum_k W(k,l) \cdot \Delta s_t(k) \cdot \quad (2)$$

Unlike MFCC, we do not use logarithm for the Mel-FB.

This paper proposes normalization with an averaged mel-spectrum $\bar{S}$ as Equation (3):

$$\Delta \hat{S}_t = \frac{\Delta S_t}{\bar{S}} \cdot \quad (3)$$

$\bar{S}$ can be calculated as Equation (4) by averaging mel-spectrums $S_t$ of the utterance. $N$ is the number of frames. For a real-time system, this can be implemented as a running update.

$$\bar{S} = \frac{1}{N}\sum_{t=1}^{N} S_t \cdot \quad (4)$$

Finally, DCT is performed to output the linear delta feature $\Delta C_t$ using Equation (5):

$$\Delta C_t = DCT\left(\Delta \hat{S}_t\right). \quad (5)$$

### *2.1.1. Justification*

The delta MFCC is based on the differences in the logarithmic domain as shown in Equation (6):

$$\Delta(MFCC)_t = \Delta\left(DCT\left(\log(S_t)\right)\right) = DCT\left(\Delta\left(\log(S_t)\right)\right). \quad (6)$$

The inner term can be approximated as Equation (7):

$$\Delta\left(\log(S_t)\right) \approx \frac{\partial}{\partial S}\log(S_t) \cdot \Delta S_t = \frac{\Delta S_t}{S_t} \cdot \quad (7)$$

It should be noted that right hand sides of Equations (3) and (7) are very similar. In Equation (3), the averaged spectrum $\bar{S}$ is used instead of the instantaneous spectrum $S_t$. This means the proposed linear delta feature derived from Equation (3) has relatively smaller values than the delta MFCC derived from Equation (7) in the reverberation region, where the instantaneous spectral magnitude is relatively small.

If we introduce the channel transfer function $F$ and mel-spectrum of the source signal $X$, Equations (3) and (7) can be re-written as Equations (8) and (9), respectively, showing both implementations include channel normalization.

$$\Delta \hat{S}_t = \frac{\Delta S_t}{\bar{S}} = \frac{\Delta(FX_t)}{F\bar{X}} = \frac{\Delta X_t}{\bar{X}} \cdot \quad (8)$$

$$\Delta\left(\log(S_t)\right) \approx \frac{\Delta S_t}{S_t} = \frac{\Delta(FX_t)}{FX_t} = \frac{\Delta X_t}{X_t} \cdot \quad (9)$$

## 2.2. Linear delta-delta feature

The process for the linear delta-delta features proposed in this paper is straightforward extension of the linear delta feature discussed in Section 2.1.

Equations (10) to (13) and Figure 3 show the base process for the linear delta-delta features of $\Delta\Delta C_t$ proposed in this paper.

$$\Delta\Delta s_t = \frac{\sum_{\theta=1}^{\Theta} \theta\left(\Delta s_{t+\theta} - \Delta s_{t-\theta}\right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \cdot \quad (10)$$

$$\Delta\Delta S_t(l) = \sum_k W(k,l) \cdot \Delta\Delta s_t(k) \cdot \quad (11)$$

$$\Delta\Delta \hat{S}_t = \frac{\Delta\Delta S_t}{\bar{S}} \cdot \quad (12)$$

$$\Delta\Delta C_t = DCT\left(\Delta\Delta \hat{S}_t\right). \quad (13)$$

## 2.3. Optional logarithmic compression

For the delta feature, we use Equation (3) without logarithmic compression in our primary approach, because Equations (3) and (7) have similar forms. As an optional proposal, we can use logarithmic compression with Equation (14) replacing Equation (3). The logarithm is modified so to continuously use negative and positive values. The additional procedures including Equation (5) are the same.

$$\Delta \hat{S}_t = \begin{cases} \log\left(\frac{\Delta S_t}{\bar{S}} + 1\right) & \text{if } \Delta S_t \geq 0 \\ -\log\left(-\frac{\Delta S_t}{\bar{S}} + 1\right) & \text{otherwise} \end{cases} \cdot \quad (14)$$

Likewise, the linear delta-delta feature based on Equation (12) can be compressed with the modified logarithm as with Equation (14).

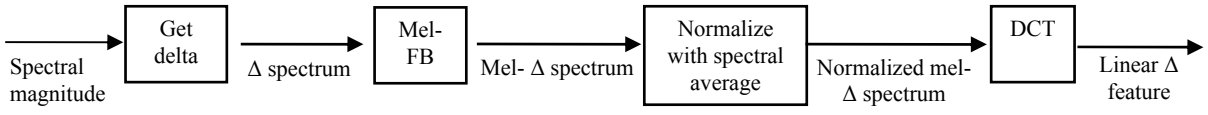


Figure 2: *Proposed linear delta feature.*

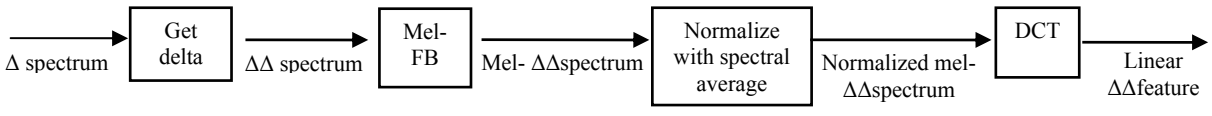


Figure 3: *Proposed linear delta-delta feature.*

## 3. Experiments

CENSREC-4 was used for the evaluation. This is a common evaluation framework for distant-talking speech recognition under reverberant environments [16][17]. It contains testing data and training data with program scripts to generate acoustic models using HTK [18].

In CENSREC-4, clean speech data was recorded with a close-talk microphone in a sound-proof chamber and impulse responses were recorded in eight kinds of rooms. Then, reverberant speech data was simulated by the convolution of the data. The reverberation times of the rooms vary from 0.05 sec to 0.75 sec.

CENSREC-4 defines two training types, clean and multi-condition. In clean training, the acoustic model is trained with non-reverberant speech data, which was prepared without convoluting impulse responses. Multi-condition training uses reverberant speech data with four kinds of room impulse responses out of eight possible kinds. The training data consists of 8,440 utterances spoken by 110 subject speakers (55 females and 55 males).

The testing data is prepared in two ways, non-reverberant and reverberant. Our main interest is in reverberant testing data, which is convolved with eight kinds of room impulse responses. Test Set A is for closed evaluations using the four room impulse responses used in the training data. Test Set B is for open evaluation using the remaining four room impulse responses. Each testing data set consists of 4,004 utterances spoken by 104 subject speakers (52 females and 52 males). The task is connected digit recognition in Japanese.

The feature parameter of the baseline system uses 39-dimension feature vectors that consist of 12-dimension MFCC, 12-dimension delta MFCC, 12-dimension delta-delta MFCC, and one dimension each for log-power, delta log-power, and delta-delta log-power. The analysis conditions were pre-emphasis, Hamming windows, 25-ms frame lengths, 10-ms frame shifts. As a baseline, CENSREC-4 does not use CMS. In our evaluation, a front-end program output various types of features to be tested, but the backend process for the acoustic models was unchanged. In the HMM, the 18 phonemes for digits have 5 states with 20 Gaussian mixtures each. Silence has 5 states and the short pause has 3 states with 36 Gaussian mixtures for each state.

### 3.1. Preliminary experiment using simplified feature combinations

In order to measure the advantages of the proposed linear delta feature, evaluations using simplified combinations of features were performed for the CENSREC-4 data. Both the training and testing data were converted to the features with the following combinations:

- 12-dimension MFCC (static only).
- 12-dimension MFCC and 12-dimension delta MFCC.
- 12-dimension MFCC and the proposed 12-dimension linear delta feature (without logarithmic compression).

No power features were used together. For the approximation, the averaged mel-spectrum $\overline{S}$ was calculated using all of the frames for each utterance file, regardless of speech or non-speech segments.

It should be noted that the static part is common to all cases and the dynamic parts including delta MFCC and the linear delta feature have their inherent channel normalizations, as we discussed in Section 2.1.1.

Table 1 shows the evaluation results. With reverberant testing data, the trial with only static the MFCC feature shows very poor performance, especially in the clean training down to 18.1%. Introducing the delta MFCC features improves to 35.3%, but that may be inadequate. Introducing the proposed linear delta feature shows impressive improvement up to 59.2%. It also shows significant improvement in multi-condition training.

Table 1. *Results of preliminary experiments using simplified feature combinations with the CENSREC-4 data. Test Set A is for matched reverberation and Test Set B is for un-matched reverberation*

| (% STRING Recognition rate) | | | mfcc 12 dim | mfcc 12 dim + delta mfcc 12 dim | mfcc 12 dim + linear delta 12 dim |
|---|---|---|---|---|---|
| Clean training | Testing without reverberation | | 87.9 | 96.1 | 95.3 |
| | Testing with reverberation | | 18.1 | 35.3 | 59.2 |
| Multi condition training | Testing with reverberation | Test Set A | 57.8 | 79.6 | 83.4 |
| | | Test Set B | 38.8 | 64.1 | 79.6 |
| | | A+B Average | 48.3 | 71.9 | 81.5 |

Table 2. *CENSREC-4 evaluation results. The baseline is the default 39-dimension features including 12-dimension delta MFCC and 12-dimension delta-delta MFCC. The Proposed 1 column replaces the delta and delta-delta with the proposed linear delta and delta-delta features. The Proposed 2 column adds the optional logarithmic compression*

| (% STRING Recognition rate) | | | Baseline | Proposed 1 | Proposed 2 |
|---|---|---|---|---|---|
| Clean training | Testing without reverberation | | 98.3 | 96.0 | 96.4 |
| | Testing with reverberation | | 65.2 | 73.1 | 73.8 |
| Multi condition training | Testing with reverberation | Test Set A | 80.7 | 82.9 | 83.2 |
| | | Test Set B | 69.6 | 82.5 | 84.1 |
| | | A+B Average | 75.2 | 82.7 | 83.7 |

### 3.2. Experiment using full feature combinations

Since the advantages of the proposed linear delta feature were presented in Section 3.1, this section describes the performance using the 39-dimension features that CENSREC-4 defines. In Table 2, “Proposed 1” is our primary proposal that replaces delta MFCC and delta-delta MFCC with the linear delta and delta-delta feature. The static part of MFCC and the power features (including delta and delta-delta) are shared with the baseline. “Proposed 2” is our secondary proposal with the optional logarithmic compression discussed in Section 2.3.

Both of the proposed methods show significant improvements over the baseline in both clean and multi-condition training with reverberant testing data. In particular, they show drastic improvements for Test Set B in multi-condition training, where the reverberant characteristics are unknown to the training model. We should note the unmatched condition is a practical assumption. In this case, the String Error Rate (SER) was reduced by 42% in “Proposed 1” and by 48% in “Proposed 2”. In contrast, the degree of improvement is relatively small in Test Set A in multi-condition training, which involves matched reverberation with high baseline performance. However, “Proposed 1” still reduced the SER by 11% and “Proposed 2” did by 13%. Overall, “Proposed 1” reduced the SER by 30% in multi-condition training, and by 23% in clean training, while “Proposed 2” did by 34% and by 25%, respectively.

At the same time, both of the proposed methods showed degradation with non-reverberant testing data. Our interpretation is that the proposed dynamic features have benefit for reverberant segments, but they have some draw back when speech power is strong. That is to say, Equation (3) was introduced to cope with reverberation, but logarithmic nature such as Equation (7) may be still necessary for strong segments. Therefore, some sort of fusion style of Equation (3) and (7) would be desired.

Comparing “Proposed 1” and “Proposed 2”, the performances were quite similar. We would recommend “Proposed 1” because of its advantage in eliminating logarithmic operations.

## 4. Conclusions

In reverberant environments, difference operations in the logarithmic spectrum domain such as the delta MFCC are affected by preceding phonemes. We have proposed a new delta feature calculated in the linear spectrum domain, an approach that quickly diminishes in reverberant segments. A similar new delta-delta feature is also introduced. Experiments using a common evaluation framework in a reverberant environment (CENSREC-4) showed significant improvements both in clean training and multi-condition training cases, and in open and closed reverberant environments with the proposed methods. However, there were degradations in non-reverberant environments.

## 5. Acknowledgements

The present study was conducted using the CENSREC-4 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

## 6. References

[1] X. Lu, M. Unoki and M. Akagi, "A robust feature extraction based on the MTF concept for speech recognition in reverberant environment," *Proc. ICSLP2006*, pp. 2546-2549, 2006.

[2] T. Nakatani, and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP2003*, vol. 1, pp. 92-95, 2003.

[3] K. Kinoshita, M. Delcroix, T. Nakatani and M. Miyoshi, "Multi-step linear prediction based speech enhancement in noisy reverberant environment," *Proc. Interspeech2007*, pp. 854-857, 2007.

[4] T. Nishikawa, H. Saruwatari and K. Shikano, “Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA,” *IEICE Trans. Fundamentals*, Vol.E86-A, No.4, pp. 846-858, 2003.

[5] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition," *Proc. ICASSP '99*, vol. 1, pp. 449-452, 1999.

[6] M. Delcroix, T. Hikichi and M. Miyoshi, "On the use of LIME dereverberation algorithm in an acoustic environment with a noise source," *Proc. ICASSP 2006*, I, pp. 825-828, 2006.

[7] A. Baba, A. Lee, H. Saruwatari, and K. Shikano, "Speech recognition by reverberation adapted acoustic models," *Proc. ASJ*, pp. 27-28, 2002 (in Japanese).

[8] L. Couvreur, C. Couvreur and C. Ris, “A corpus-based approach for robust ASR in reverberant environments,” *Proc. ICSLP 2000*, vol. 1, pp. 397-400, 2000.

[9] K. Nishiki, S. Watanabe, T. Nishimoto, N. Ono, S. Sagayama, “A Study on Robust Speech Recognition against Unknown Reverberation Using Single Speech Model Trained under Multiple Reverberant Environments,” *IEICE Technical Report*, 108, 66, pp.43-48, May., 2008 (in japanese).

[10] M. J. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," *Cambridge University Engineering Department Technical Report*, 1997.

[11] M.J.F. Gales and S.J. Young, "Robust speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, pp. 352-359, 1996.

[12] C. J. Leggetter and P. C.Woodland, "Maximum likelihood linear regression for speaker adaption of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, pp. 171-185, 1995.

[13] M. L. Shire and B. Y. Chen, "Data-driven RASTA filters in reverberation," *Proc. ICASSP 2000*, Vol. 3, pp. 1627-1630, 2000.

[14] T. Takiguchi and Y. Ariki, "Robust feature extraction using Kernel PCA," *Proc. ICASSP 2006*, pp. 509-512, 2006.

[15] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. ASSP-34, No. 1, pp. 52-59, 1986.

[16] M. Nakayama et al., "CENSREC-4: development of evaluation framework for distant-talking speech recognition under reverberant environments," *Proc. Interspeech 2008*, pp. 968-971, 2008.

[17] T. Nishiura et al., “Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments - Newest Part of the CENSREC Series -,” *Proc. LREC '08*, 2008.

[18] http://htk.eng.cam.ac.uk/