WEIGHTED LIKELIHOOD RATIO (WLR) HIDDEN MARKOV MODEL FOR NOISY SPEECH RECOGNITION

Chao Huang[1], Yingchun Huang[1,2], Frank Soong[1] and Jianlai Zhou[1]*

[1]Microsoft Research Asia
[2]Institute of Electronics, Chinese Academy of Sciences
{chaoh, jlzhou,frankkps}@microsoft.com; p-ychuang@msrchina.research.microsoft.com

ABSTRACT

In this paper we present a weighted likelihood ratio (WLR) based Hidden Markov Model and apply it to speech recognition in noise. The WLR measure emphasizes spectral peaks than valleys in comparing two given speech spectra. The measure is more consistent with human perception of speech formants where natural resonances of vocal track are and tends to be more robust to broad-band noise interferences than other measures. A complete HMM framework of this measure is derived and a mixture of exponential kernels is used to model the output probability density function. The new WLR-HMM is tested on the Aurora2 connected digits database in noise. It shows more robust performance than the MFCC trained GMM baseline system. When combined with the dynamic cepstral features, the multiple-stream WLR-HMM shows a 39% relative improvement over the baseline system.

1. INTRODUCTION

As speech recognition is transferred from the laboratory to marketplace, robust recognition is becoming increasingly important and critical. Among all kinds of the variabilities and mismatches such as speakers, accents and channels, background noise is one of the hardest problems and the most usual cases in real speech applications we face. It is why noise robust attracting lots of researchers' interest.

There are all kinds of algorithms to deal with it. Missing feature algorithm tries to extract features which are more invariant or insensitive to noises in spectro-temporal regions as [5]. Ealey and etc. recover the underlying speech through making an improved estimate of noise spectrum by making fully use of the harmonic structure of the voiced speech spectrum [6]. Others include adding weights for different front-end according to their relative sensitiveness to noises [3].

Weighted Likelihood Ratio (WLR) was first proposed in 1984 by Sugiyama [2] as a distortion measure when comparing two given speech spectra. More emphasis has been put to the peak part of the spectrum during the measuring. It is not only consistent with human perception, but also accordance with the fact the peak (formant) plays a more important role during the recognition. Especially it should be noted that peak part is much less polluted by noises. It is successfully used for vowel classification and isolated word recognition based on DP.

In this paper, we introduce the concept of the peak-weighting and extend WLR in following aspects:

1. Replacing LPCC in [2] with MFCC and deriving MFCC based WLR since MFCC is widely confirmed by researchers that it is more effective than LPCC in recognition.
2. Instead of DP and codebooks used in [2], we derive WLR-based HMM and make it seamlessly combined with state-of-the-art HMM framework.
3. According to [3], dynamic spectral information of MFCC are more robust to noise than static one, we combine WLR-HMM with conventional HMM based on dynamic-MFCC through a two-stream HMM.

In Section 2, a complete HMM framework based on WLR measure is derived. In Section 3, we evaluate the proposed 2-stream WLR-HMM on the Aurora2 database and analyze the results. Conclusion is given in Section 4.

2. WLR-HMM

In this section, we will first review the basic idea of WLR and show why it is more suitable for speech recognition in noise. Then we will describe the procedure of building MFCC-based WLR and then derive the WLR-HMM framework. Finally we will introduce a two-stream HMM

[*] Join this work as a visiting student at MSRA

system that combining WLR-HMM and conventional HMM based on dynamic MFCC.

2.1. WLR

The WLR measure emphasizes the spectral peaks than valleys in comparing two given speech spectra. This measure is more consistent with human perception of speech formants where natural resonances of vocal track are and tends to be more robust to noise interferences than other measures where no emphasis is placed on the spectral peaks. Since in terms of local (in frequency) SNR, the peak parts of spectrum are less polluted by noises. It can be illustrated by Figure 1, where topper part is the linear power spectra and the bottom part is the corresponding log spectra. Blue (real), red (dash) and green (bar) represent clean, noisy (SNR=5dB) spectra and their differences respectively. Usual cepstrum distortion is shown by the green (bar) of bottom figure and the main differences are due to the valley parts which are tender to be affected by noises. According to WLR, which weights log spectral difference with the linear spectral difference, the distortion becomes much less. It is what we expected. In other word, difference in valley parts which are less reliable is de-emphasized and that in peak parts which are more reliable is emphasized.

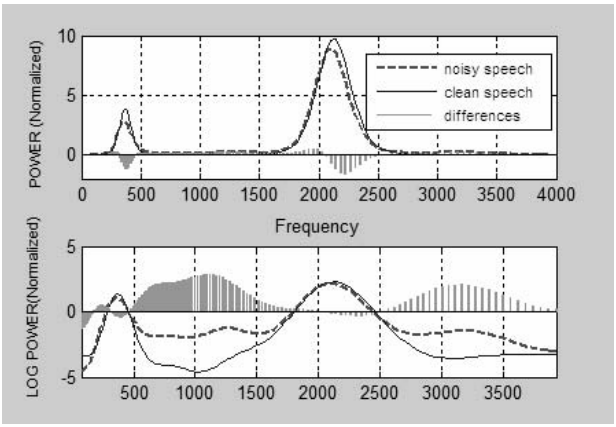

Figure 1: Illustration of WLR

WLR can be formulated by (2.1) where in integrands, $\log S_t(w) - \log S_r(w)$ is the difference between two log spectra: test spectrum $\log S_t(w)$ and reference spectrum $\log S_r(w)$. $S_t(w) - S_r(w)$, the difference between the corresponding linear spectra, is used as the weighting function.

$$d_{wlr}(\log S_t(w), \log(S_r(w)) = \int_{-\pi}^{\pi} [(S_t(w) - S_r(w)][(\log S_t(w) - \log S_r(w)] \frac{d\lambda}{2\pi} \quad (2.1)$$

According to Passeval's theorem, WLR spectral distortion can be re-formulated as (2.2)

$$d_{wlr}(\log S_t(w), \log(S_r(w)) = \sum_{i=-\infty}^{+\infty} [(r_t(i) - r_r(i))(c_t(i) - c_r(i))] \quad (2.2)$$

Here $r_t(i)$ and $c_t(i)$ are the autocorrelation and cepstral coefficients of the test spectrum, respectively. Same is true for reference spectrum. It should be noted that weighting function should satisfy (2.3). In other word, the 0^{th} coefficients of $r_t(i)$ and $r_r(i)$ are constrained to unity power, or 1.

$$\int_{-\pi}^{\pi} S_t(w) \frac{d\lambda}{2\pi} = 1 \qquad \int_{-\pi}^{\pi} S_r(w) \frac{d\lambda}{2\pi} = 1 \quad (2.3)$$

2.2. MFCC based WLR

It is reported Mel-frequency cepstrum coefficients (MFCC) are more effective and more robust for speech recognition

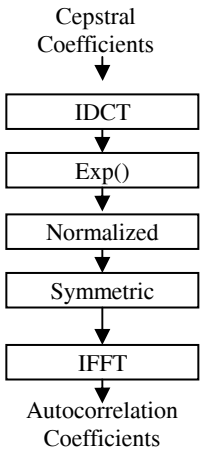

Figure 2: Block diagram of extracting the weighting function (Autocorrelation coefficients) from MFCC

than linear prediction cepstral coefficients (LPCC) and widely used by current state-of-the-art recognizers. Therefore, we have implemented MFCC-derived WLR: Cepstra used in (2.2) is MFCC instead of LPCC. Here the Arithmetical mean of MFCC is used to approximate the centroids of the WLR-based measure although it is not exactly true by strict definition of the measure. Given the MFCC, we can derive the corresponding weighting function: autocorrelation coefficients by following these steps described in Figure 2.

2.3. WLR-HMM

It is obvious to verify the WLR distortion values are non-negative from (2.1) since log function is monotonic and thus the difference of linear spectra has the same +/- sign as the corresponding parts of log spectra in the integrand, and therefore the integrand is semi-positive. A mixture of exponential kernels can be used to model the output probability density function (pdf) as shown in (2.4) and as a whole it is called WLR-HMM.

$$b_j(o_t) = \sum_{k=1}^{M} w_{jk}\beta_{jk} \exp(-\beta_{jk} * d_{wlr}(o_{t,} u_{jk})) \qquad (2.4)$$

Here, $o_{t,}$ is the observation vector consisting of $r_t(i)$ and $c_t(i)$, and u_{jk} is the mean vector and β_{jk} is the inverse mean of the WLR distortion of the *j*-th state and *k*-th component. And w_{jk} is the weighing coefficient of *k*-th component for *j*-th state. In practice, pdf can also be realized as (2.5) form.

$$b_j(o_t) = \max_{k=1,2..M}\{w_{jk}\beta_{jk} \exp(-\beta_{jk} * d_{wlr}(o_{t,} u_{jk}))\} \qquad (2.5)$$

The auxiliary Q-function for WLR-HMM density can be written as:

$$Q(\bar{\theta}, \theta) = \sum_{q} P(O, q \mid \theta) \cdot \log P(O, q \mid \bar{\theta}) \qquad (2.6)$$

By taking the partial derivative of right side of (2.5) regard to each parameter and let them equal to 0, the updated β_{jk} , centroids and kernel weights are derived and given as:

$$\bar{\beta}_{jk} = \frac{\sum_{t=1}^{T} \psi_{jk}(t)}{\sum_{t=1}^{T} \psi_{jk}(t) \cdot d_{wlr}(o_t, u_{jk})} \qquad (2.7)$$

$$\bar{u}_{jk} = \frac{\sum_{t=1}^{T} \psi_{jk}(t) \cdot o_t}{\sum_{t=1}^{T} \psi_{jk}(t)} \qquad (2.8)$$

$$\bar{w}_{jk} = \frac{\sum_{t=1}^{T} \psi_{jk}(t)}{\sum_{k=1}^{M}\sum_{t=1}^{T} \psi_{jk}(t)} \qquad (2.9)$$

Where, $\psi_{jk}(t)$ is an indicator function which is 1 if o_t is associated with the *k*-th component of the *j*-th state and is zero otherwise.

2.4. 2-stream WLR-HMM

According to Yang's work [3], dynamic cepstral features play more important roles especially for noisy speech recognition. As we know, WLR-HMM can help improve the noise robustness of static MFCC by more robust distortion measure as it will be shown by experiments in the next section. The simple way to improve the performance further is to merge them together. It can be formulated by (2.10) which integrate them by two-stream in the level of computing the likelihood scores. Weighting coefficients γ_1 and γ_2 are used to reflect the relative importance and normalize the different dynamic ranges of scores from these two streams,

$$b_j(o_t) = [\sum_{k=1}^{M} w_{jk}\beta_{jk} \exp(-\beta_{jk} * d_{wlr}(o_t^{wlr}, u_{jk}^{wlr}))]^{\gamma_1} * [\sum_{k=1}^{M} c_{jk} N(o_t^d; u_{jk}^d, \Sigma_{jk}^d)]^{\gamma_2} \qquad (2.10)$$

The weighting coefficients can be learnt through limited development set and they are tuned experimentally now.

3. RESULTS

3.1. Experimental Setups

Throughout the paper, Aurora2 database are used for the evaluations. Only clean data are used for both baseline and WLR-HMM model training. Testing set include A, B and C. During the training and testing, only male data are used for both baseline and WLR-HMM model. For the baseline, model training is full consistent with the standard recipe described in [1], including the number of states per word, components per state, iteration procedure and so on.

For WLR-HMM, only 13 order of MFCC (including C0) are used. No cepstral lifter is used when extracting MFCC to satisfy (2.2). C0 is not practically used for WLR-HMM since corresponding $r_t(0) = r_r(0) = 1$.

All the number shown in the following tables are the average of the accuracy from 0dB to 20dB except that stated explicitly.

3.2. WLR-HMM based only Static Features

First we evaluated WLR-HMM performance based on static cepstral features of MFCC only. Configuration MFCC-S in Table 1 means only MFCC-E front-end (13d) is used in standard Aurora2 training procedure. WLR-Init in Table 1 means the initial WLR-HMM is directly computed from the model of MFCC-S. With several more iterations of WLR-HMM training, we obtained final WLR-HMM model sets (WLR-HMM at Table 1). Table 2 shows the comparison results.

Accuracy	Subway	Babble	Car	Exhibition	Overall
MFCC-S	36.37	41.77	39.59	38.61	39.08
WLR-Init	47.40	42.09	54.86	53.73	49.52
WLR-HMM	53.80	50.89	58.78	55.86	54.83
Relative Improvement	27%	16%	32%	28%	26%

Table 1: WLR-HMM vs. conventional HMM based on static MFCC only (Testing set A with clean training set)

It shows that only static feature, WLR-HMM can improve the performance. The comparatively less improvement on babble noise is probably due to speech-like characteristic of this kind of noise which is also emphasized with speech, especially for pure noise segments like the beginning and the ending of utterances.

3.3. WLR-HMM with Dynamic MFCC

Based on studies in [3], dynamic cepstral feature is more robust in noise than static one. 2-stream WLR-HMM is then built to combine WLR-HMM and conventional dynamic MFCC based HMM as described in section 2.4. Weighting coefficients of the two streams are tuned experimentally.

Accuracy	Set A	Set B	Set C	Overall
MFCC(39d)	61.38	57.50	68.19	**62.36**
WLR-HMM	76.40	78.44	75.77	**76.87**
Relative Improvement	**39%**	**49%**	**24%**	**39%**

Table 2: Summary of comparisons between 2-stream WLR-HMM and baseline (39d MFCC):

Overall results are summarized in Table 2 and more details are given in Table 3. 2-stream WLR-HMM greatly improves the noise robustness although its performance still varies from noise to noise.

It is very interesting that among the 8 noises, WLR-HMM achieved relatively smaller improvements on subway, exhibition and street noises. While we investigated the corresponding noise spectra given by [1], they show peak-like characteristics like speech. Especially for subway and exhibition noises, they are emphasized similarly as speech spectral peaks by WLR-HMM at the same time.

4. CONCLUSIONS AND DISCUSSIONS

A complete HMM framework based on WLR measure, called WLR-HMM is proposed in the paper. After combining with dynamic cepstral features, multiple-stream WLR-HMM show 39% relative improvement over the baseline system. As a measure that is more consistent with human perception of speech formants, WLR-HMM shows experimentally more robust recognition performance than the standard MFCC baseline system in noise. No noise estimation is needed for the WLR-HMM.

WLR-HMM are not as effective for lower amplitude, unvoiced sound like fricatives or broadband noises where no distinctive formant-like spectral peaks exist. A further rescoring with WLR-HMM on the lattice decoded with the baseline may be a better alternative.

5. REFERENCES

[1] Hams-Gunter Hirsch and David Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition System under Noisy Conditions", Proc. of ISCA-ITRW ASR2000, pp.181-188, Sept. 2000.

[2] Masahide SUGIYAMA, "LPC Spectral Matching Measures for Speech Recognition", Ph.D. dissertation, Tohoku Univ., Aug.1984.

[3] Chen Yang, Frank S. Soong and Tan Lee, "Static and sDynamic Spectral Features: Their Noise Robustness and Optimal Weights for ASR", ICASSP2005, Philadelphia, USA.

[4] The HTK Toolkit: http://htk.eng.cam.ac.uk.

[5] Martin Cooke, Phil Green, Ljubomir Josifovski and Ascension Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data", Speech Communication 34 (2001) pp. 267-285.

[6] Douglas Ealey, Holly Kelleher and David Pearce, "Harmonic Tunneling: Tracking Non-station Noises during Speech", Eurospech'2001, Demark.

Aurora 2 Clean Training - Results													
	A					B					C		
	Subway	Babble	Car	Exhibition	Aver.	Restaurant	Street	Airport	Station	Aver.	Subway MIRS	Street MIRS	Aver.
SNR-5	16.61	15.56	10.71	11.31	**13.55**	12.41	15.00	18.00	13.95	**14.84**	15.95	17.17	**16.56**
SNR0	35.91	37.71	30.68	32.43	**34.18**	37.71	39.39	45.16	37.52	**39.95**	35.85	37.21	**36.53**
SNR5	66.97	71.69	66.33	64.49	**67.37**	69.84	71.50	73.98	68.95	**71.07**	62.77	67.58	**65.18**
SNR10	86.99	89.48	88.52	85.98	**87.74**	87.83	88.80	90.55	89.82	**89.25**	83.03	87.55	**85.29**
SNR15	94.54	95.40	96.29	94.78	**95.25**	93.76	96.27	95.16	95.10	**95.07**	93.65	95.02	**94.34**
SNR20	97.18	97.82	97.97	96.92	**97.47**	97.18	97.26	96.17	96.92	**96.88**	97.18	97.82	**97.50**
Clean	98.20	98.32	98.39	97.67	**98.15**	98.20	98.32	98.39	97.67	**98.15**	98.14	98.26	**98.20**
Aver.	76.32	78.42	75.96	74.92	**76.40**	77.26	78.64	80.20	77.66	**78.44**	74.50	77.04	**75.77**
Baseline	70.29	49.37	59.68	66.17	**61.38**	54.31	65.75	53.2	56.75	**57.50**	65.92	70.45	**68.19**
RR-WER	**20%**	**57%**	**40%**	**26%**	**39%**	**50%**	**38%**	**58%**	**48%**	**49%**	**25%**	**22%**	**24%**

Table 3: Results from two-stream WLR-HMM on Aurora2 with clean training (RR-WER in last row means the relative reduction of WER of 2 stream WLR-HMM compared with the baseline).