ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

DERIVING MFCC PARAMETERS FROM THE DYNAMIC SPECTRUM FOR ROBUST SPEECH RECOGNITION

Nengheng Zheng, Xia Li

College of Information Engineering
Shenzhen University
Nanhai Ave 3688, Shenzhen, Guangdong, 518060

Houwei Cao, Tan Lee, P. C. Ching

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, NT, Hong Kong

ABSTRACT

State-of-the-art automatic speech recognition systems typically adopt the feature set containing Mel-frequency cepstral coefficients (MFCC) and their time derivatives. The noise vulnerability of MFCC significantly degrades the recognition performance of such systems in noisy conditions. This paper describes a noise-robust feature extraction method. A set of new MFCC features is derived from the dynamic spectrum instead of the static spectrum as in the conventional MFCC feature extraction. It is shown that the dynamic spectrum preserves the spectral envelope information and, at the same time, is more noise resistant than the static spectrum. Experiments on Aurora 2 database show the noise robustness of the proposed features and it is preferable to replace MFCC with the new features in the state-of-the-art feature set.

Index Terms— Speech recognition, dynamic spectrum, noise robustness, MFCC

1. INTRODUCTION

Cepstral coefficients have been widely accepted as the representative acoustic features in state-of-the-art speech recognition systems. The cepstral coefficients, by definition, are obtained by inverse Fourier transform the logarithm of the magnitude spectrum of speech. The well known MFCC features can attain high recognition accuracy in controlled environments [1]. The performance in real-world applications, however, tends to degrade significantly because of the mismatch between distorted input speech and the pre-trained acoustic models. The most common types of distortion include additive noise contamination and convolutive noise distortion caused by microphones and transmission channels. This study focuses on extracting robust features to alleviate the additive noise problem.

Approaches to robust feature extraction include: applying noise suppression techniques to increase signal-to-noise ratio (SNR) of input speech [2], cepstral mean normalization [3], use of dynamic cepstral coefficients [4] and vector Taylor series [5], etc. The computation of dynamic cepstral coefficients, e.g., ΔMFCC, is equivalent to a spectral filtering process in the log-spectral domain. Theoretically it serves to remove convolutive noise. Yang [6] demonstrated that ΔMFCC is more robust to additive noise than MFCC. On the other hand, spectral filtering in the linear spectral domain, i.e., magnitude or power spectral domain, can reduce additive noise, if speech and noise are uncorrelated. Feature extraction with spectral filtering in linear spectral domain was reported in Hirsch [7] and Xu [8]. Hermansky proposed the RASTA technique, in which the noisy speech spectrum is filtered either in the linear spectral domain or in the log-spectral domain, depending on the SNR estimated [9].

To achieve the best recognition performance, state-of-the-art systems typically combine MFCC and ΔMFCC (and sometimes also $\Delta\Delta$MFCC) to form the feature vector. Given that MFCC is noise vulnerable, the combined feature vector is not noise resistant, especially when the noise level is high. It is desirable to derive a set of new features that can attain as good performance as MFCC in clean conditions and, at the same time, is robust to noise distortion.

In this study, we describe and evaluate a feature extraction technique that generates noise-resistant cepstral coefficients. Unlike the computation of conventional MFCC, which is derived from the static spectrum, we propose to compute the cepstral coefficients from dynamic spectrum (the dynamic spectrum in linear spectral domain). It is observed that the dynamic spectrum has similar envelope to the static spectrum and thus contains useful information for phonetics classification. Furthermore, the dynamic spectrum is expected to be more noise resistant than the static spectrum, provided that noise and speech are additive and that noise changes slowly as compared to speech. The effectiveness of the proposed features is evaluated in a set of experiments carried out on the Aurora 2 database.

2. DYNAMIC SPECTRUM OF SPEECH

Let $S(\omega)$, $N(\omega)$ and $S_N(\omega)$ be the frequency spectra of speech, noise and noise corrupted speech signals, respectively. Assuming that noise is additive, we have,

$$S_N(\omega) = S(\omega) + N(\omega) \tag{1}$$

The power spectral density of noisy speech can be computed as,

$$\begin{aligned} |S_N(\omega)|^2 &= [S(\omega)+N(\omega)][S^*(\omega)+N^*(\omega)] \\ &= |S(\omega)|^2 + |N(\omega)|^2 + 2|S(\omega)||N(\omega)|\cos(\alpha-\beta) \end{aligned} \tag{2}$$

where α and β are the phase angles of $S(\omega)$ and $N(\omega)$, respectively. Both α and β are not known. There are two special cases as described below:

Case 1: $\alpha - \beta = 0$, we have

$$|S_N(\omega)| = |S(\omega)| + |N(\omega)| \tag{3}$$

Case 2: $\alpha - \beta = \pi/2$, we have

$$|S_N(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 \tag{4}$$

This work is jointly supported by a research grant **Project 200871** awarded by Shenzhen University R/D Fund and a research grant from the Shun Hing Institute of Advanced Engineering, the Chinese University of Hong Kong.The first author performed part of this work while at The Chinese University of Hong Kong.

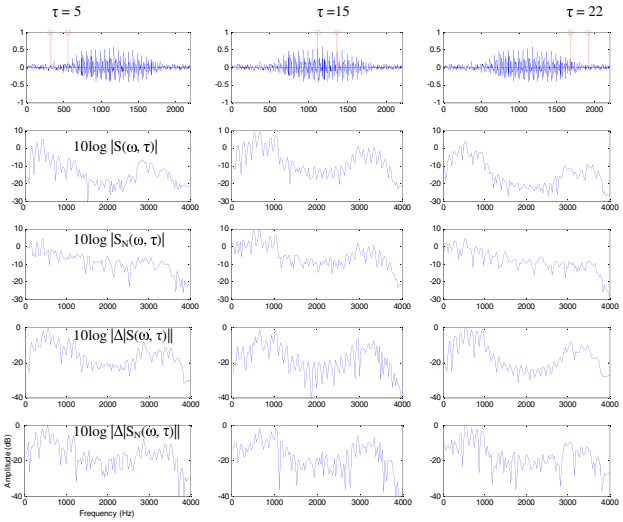

Fig. 1. The static and dynamic spectra of clean and noisy speeches. From top to bottom are the speech waveforms, static spectra of clean and noisy speeches, and the dynamic spectra of clean and noisy speeches, respectively.

Assuming that the noise signal is stationary, for the τ-th frame, (3) can be rewritten as

$$|S_N(\omega,\tau)| = |S(\omega,\tau)| + |N(\omega)| \tag{5}$$

The dynamic magnitude spectrum of the τ-th frame is computed using a regression formula, as commonly adopted in calculating the dynamic cepstral coefficients [4], i.e.,

$$\Delta|S(\omega,\tau)| = \frac{\sum_{k=-K}^{k=K} k|S(\omega,\tau+k)|}{2\sum_{k=-K}^{k=K} k^2} \tag{6}$$

where K defines the window size within which the dynamic information is concerned. From (5) and (6), we have

$$\begin{aligned}\Delta|S_N(\omega,\tau)| &= \frac{\sum_{k=-K}^{k=K} k\{|S(\omega,\tau+k)| + |N(\omega)|\}}{2\sum_{k=-K}^{k=K} k^2} \\ &= \frac{\sum_{k=-K}^{k=K} k|S(\omega,\tau+k)|}{2\sum_{k=-K}^{k=K} k^2}\end{aligned} \tag{7}$$

That is, $\Delta|S_N(\omega,\tau)| = \Delta|S(\omega,\tau)|$. Similarly, for Case 2, we have the same relation for dynamic power spectra of noisy and clean speech: $\Delta|S_N(\omega,\tau)|^2 = \Delta|S(\omega,\tau)|^2$.

Although the two assumptions in Case 1 and 2 are usually not exactly the real case, it is found that the dynamic spectrum of the noisy speech does approximate the clean one well, under the stationary noise assumption. Fig. 1 shows the logarithm of the static and dynamic magnitude spectra of three clean speech frames in an utterance and their noisy counterparts. The SNR is 10 dB in this case. It can be seen that, for clean speech, the envelopes of the static spectra and the dynamic spectra are very similar. Thus the dynamic spectrum is expected to be as effective as the static spectrum in speech recognition. For noisy speech, it is clear that the dynamic spectra are relatively less affected by the additive noise than the static ones, especially when the SNR is low (e.g., $\tau = 5$ and $\tau = 22$ in Fig. 1). We have similar observations on dynamic power spectra.

3. NOISE-ROBUST PARAMETERS: MFCC$_{\Delta S}$

3.1. Feature extraction

To generate cepstral coefficients from the dynamic spectrum, we follow the same procedure as conventional MFCC feature extraction. The only difference is that the subband magnitude spectrum $S_B(B_i,\tau)$ is replaced by its time derivative $\Delta S_B(B_i,\tau)$.

1) Short-time Fourier transform is applied every 10 ms with a 30 ms Hamming window.
2) The spectrum is warped with a Mel-scale filter bank that consists of 26 filters. The magnitude of each filter output $S_B(B_i,\tau)$ is calculated.
3) Calculate $\Delta S_B(B_i,\tau)$ by (6).
4) Discrete cosine transform is applied to $\log|\Delta S_B(B_i,\tau)|$. The first 13 cepstral coefficients constitute a new feature vector, noted as MFCC$_{\Delta S}$.

3.2. Relation among MFCC, MFCC$_{\Delta S}$ and ΔMFCC

Cepstrum is defined as the inverse Fourier transform of the log-spectrum, i.e.,

$$c(m,\tau) = \int_0^{2\pi} \log|S(\omega,\tau)| e^{j\omega m} d\omega \tag{8}$$

And we have the following relations for $\Delta c(m,\tau)$ and $c_{\Delta S}(m,\tau)$

$$\Delta c(m,\tau) = \int_0^{2\pi} \Delta\log|S(\omega,\tau)| e^{j\omega m} d\omega \tag{9}$$

$$c_{\Delta S}(m,\tau) = \int_0^{2\pi} \log|\Delta|S(\omega,\tau)|| e^{j\omega m} d\omega \tag{10}$$

That is, $c(m,\tau)$, $\Delta c(m,\tau)$ and $c_{\Delta S}(m,\tau)$ are derived from the static spectrum, the dynamic log-spectrum and the log dynamic-spectrum, respectively. Similarly, MFCC, MFCC$_{\Delta S}$ and ΔMFCC characterize the subband spectra and the subband spectral dynamics in the magnitude and the log-magnitude spectral domains, respectively. Yang [6] demonstrated that ΔMFCC is more robust to noise than MFCC. According to the analysis in Section 2, MFCC$_{\Delta S}$ is expected to be more noise robust than MFCC.

Fig. 2 compares the speech recognition accuracy obtained with MFCC, MFCC$_{\Delta S}$ and ΔMFCC, respectively, on the Aurora 2 database. Details on the experimental setup will be introduced in section 4. From Fig. 2, it is clear that for noisy speech, the proposed MFCC$_{\Delta S}$ outperforms the MFCC (e.g., 18% absolute accuracy improvement at SNR=15 dB). For clean speech, the accuracy given by MFCC$_{\Delta S}$ is only slightly lower than MFCC. ΔMFCC outperforms MFCC and MFCC$_{\Delta S}$ for both clean and noisy speech.

3.3. Comparison to RASTA filtering

RASTA processing has been demonstrated to be a very successful technique for robust feature extraction for ASR [9]. The RASTA processing essentially performs a filtering of the speech spectrum and the transfer function of the RASTA filter is given by

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}} \tag{11}$$

Similarly, the calculation of the dynamic spectrum as (6) (from now on, noted as Delta processing), wiht $K = 2$, corresponds to a filtering of the spectrum with the transfer function

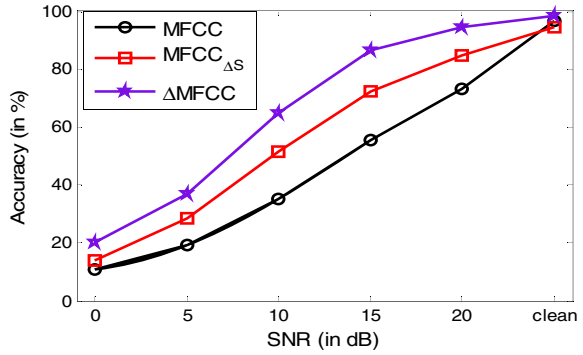

Fig. 2. Recognition accuracy by MFCC, MFCC$_{\Delta S}$ and ΔMFCC.

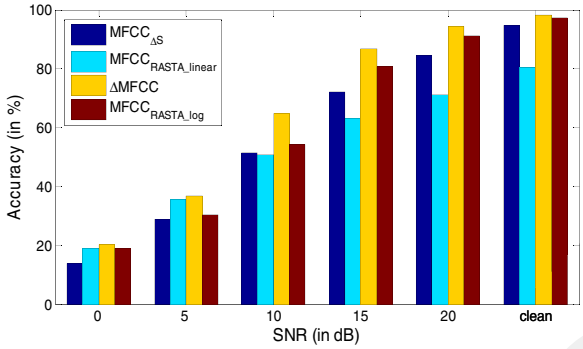

Fig. 3. Comparison of the Delta processing and RASTA processing.

$$D(z) = 0.1z^2 * (2 + z^{-1} - z^{-3} - 2z^{-4}) \quad (12)$$

Since zeros at the origin do not affect the spectrum magnitude, the only difference between RASTA and Delta filters is the extra pole in (11), which results in a different pass band for the speech modulation spectrum [10]. Hermansky proposed a lin-log RASTA technique which filters the linear spectrum at low SNR frequency bins where additive noise is dominant; and filters the log-spectrum at high SNR bins where convolutive noise is dominant [9]. In this paper, we compare the effectiveness of the two filters $H(z)$ and $D(z)$ in generating the additive noise robust MFCC features, when the filtering is on the linear and log-spectra, respectively. Fig.3 shows that in linear spectrum, RASTA filtering results in higher recognition accuracy than Delta processing at very low SNR cases, i.e., 0 and 5 dB; as SNR increases, the superiority of Delta processing comes to be more noticeable. For the log-spectrum, on the other hand, Delta processing always outperforms RASTA for all SNR cases.

For simplicity, in the following experiments, we only compare MFCC with MFCC$_{\Delta S}$, since RASTA processing only has superiority over Delta processing at 0 and 5 dB SNR in the linear spectrum.

4. EXPERIMENTS

4.1. The Aurora 2 database

The Aurora 2 database has been widely used for development and evaluation of noise-robust speech recognition systems [11]. The database contains both clean and noisy speech data. The clean speech is the 8 kHz downsampled TIDIGIT utterances. The noisy speech data were obtained by artificially adding noise to the clean data, with different types of noise at various SNRs.

The database contains both clean and multi-condition training data. The clean training data consist of 8440 utterances. The multi-condition training data consist of 20 subsets, 422 utterances per subset. Four of the subsets contain clean data. The other 16 subsets contain noise contaminated data. The noise types include: subway, babble, car and exhibition, and the SNRs are 5, 10, 15 and 20 dB.

There are three sets of test data. Set A has the same noise types as the multi-condition training data. Set B was contaminated by four different types of noise: restaurant, street, airport, and train station. Set C was first corrupted with a channel distortion and subsequently contaminated by the subway and street noises [11]. The SNRs in all sets are from -5 to 20 dB, plus the clean condition.

4.2. Experimental setup

The noise robustness of MFCC, MFCC$_{\Delta S}$ and ΔMFCC are compared over the Aurora 2 database. In addition, systems with combined features: MFCC+ΔMFCC, MFCC$_{\Delta S}$+ΔMFCC, and that plus the $\Delta\Delta$MFCC are also evaluated to demonstrate the superiority of MFCC$_{\Delta S}$ over MFCC in noise-robust speech recognition.

For each system, as in [11], 11 whole-word HMMs are trained to model the English digits, "zero" to"nine", plus "oh". Each model has 16 states without state skipping. Each state's output pdf is a mixture of three Gaussians with diagonal covariance matrices. There are also a three-state "silence" model and a single-state "short pause" model.

4.3. Recognition performance with clean training

The recognition performances with clean training are presented in Fig. 4, which compares the word accuracy of different systems at various SNRs and the clean condition. In each test set, for a specific SNR, the average accuracy of all noise types is given. It is clear that for noisy speech, regardless of the noise types and SNRs, MFCC$_{\Delta S}$ outperforms MFCC. For clean speech, the accuracy given by MFCC$_{\Delta S}$ is only slightly lower than that by MFCC. ΔMFCC outperforms both MFCC and MFCC$_{\Delta S}$ at all conditions. For the two systems with combined features, it should be noted that although combining MFCC and ΔMFCC results in a slightly improved performance, compared with ΔMFCC only, for clean speech, including MFCC usually degrades the performance for noisy speech due to the noise vulnerability of MFCC. On the other hand, combining MFCC$_{\Delta S}$ and ΔMFCC outperforms ΔMFCC for Set A and C, especially at low SNRs. For Set B, including MFCC$_{\Delta S}$ slightly degrades the performances, but still outperforms that combining MFCC and ΔMFCC. For clean speech, the two combining systems give nearly the same accuracy for all sets. In a word, it is usually beneficial to replace MFCC with MFCC$_{\Delta S}$ for robust speech recognition.

4.4. Recognition performance with multi-condition training

Fig. 5 illustrates the performances with multi-condition training. As expected, all systems show noticeable performance improvements compared with clean training. ΔMFCC outperforms both MFCC and MFCC$_{\Delta S}$ for all noisy test data. Unlike in the clean training cases, MFCC and MFCC$_{\Delta S}$ show different noise robustness in the three sets. For Set A, MFCC outperforms MFCC$_{\Delta S}$. For Set B, comparative performances are obtained. For Set C, MFCC$_{\Delta S}$ outperforms MFCC. As described in Section 4.1, from Set A to C, there are increasing degree of mismatch between the multi-condition training data and the test data. The noise types in Set A are the same as those in multi-condition training data. For Set B, the noise types differ from the training data. Further mismatches exist for Set C

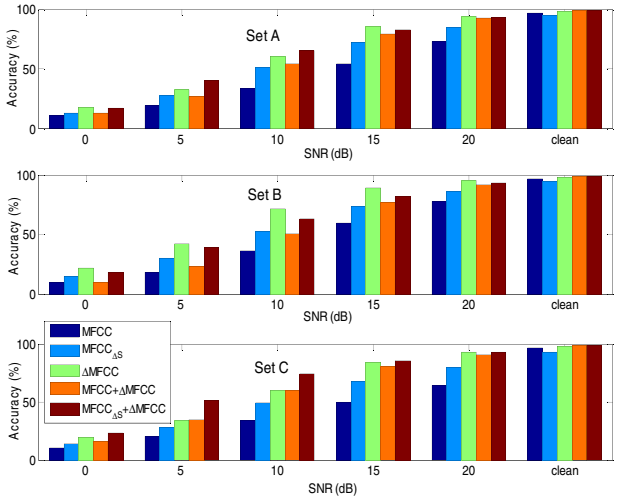

Fig. 4. Word accuracy with clean training.

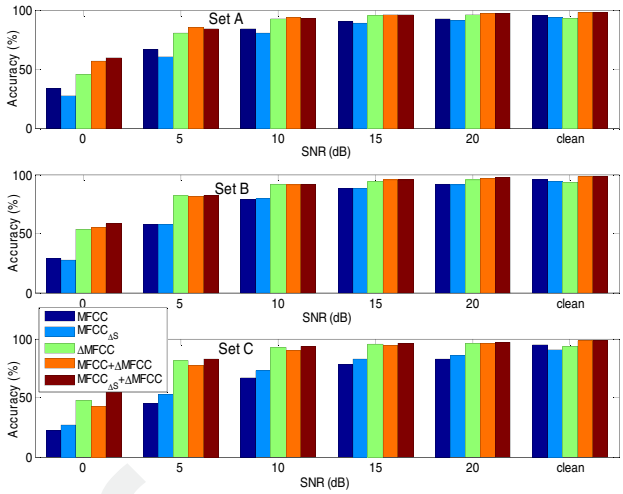

Fig. 5. Word accuracy with multi-condition training.

since the channel distortion is introduced. The results clearly tell the superiority of MFCC$_{\Delta S}$ over MFCC in speech recognition with mismatched training and test conditions. Fig. 5 shows that, in most cases, it is preferable to combining MFCC$_{\Delta S}$ and ΔMFCC, instead of combining MFCC and ΔMFCC, especially when there is significant mismatch as in Set C tests.

4.5. Including the ΔΔMFCC

Besides ΔMFCC, the 2nd-order dynamic coefficients ΔΔMFCC are also included as feature parameters in some recognition systems. Table 1 compares the average word accuracy for the three test sets given by two feature sets: MFCC+ΔMFCC +ΔΔMFCC and MFCC$_{\Delta S}$+ΔMFCC +ΔΔMFCC. As illustrated, replacing MFCC with MFCC$_{\Delta S}$ results in improved recognition performance for all three test sets.

Table 1. Average word accuracy (in %) for Test Set A-C by two feature sets (MFCC+ΔMFCC +ΔΔMFCC / MFCC$_{\Delta S}$+ΔMFCC +ΔΔMFCC)

Training model	Test Set A	Test Set B	Test Set C
Clean	67.5 / 69.6	64.1 / 69.0	73.6 / 76.3
Multi-condition	88.3 / 88.7	88.1 / 88.4	86.8 / 87.8

5. CONCLUSIONS

The use of combined MFCC and ΔMFCC as feature parameters have been widely adopted in speech recognition systems. However, the noise vulnerability of MFCC makes the combined features not robust to noise contamination. we propose to use the MFCC$_{\Delta S}$, which is the MFCC parameters derived from the dynamic spectrum, to replace the conventional MFCC. We demonstrate that the dynamic spectrum preserves the spectral envelope information and, at the same time, is more noise resistant than the static spectrum, especially for low SNR speech. Speech recognition experiments on the Aurora 2 database show that MFCC$_{\Delta S}$ significantly outperform MFCC for noisy speech and performs comparatively to MFCC for clean speech. In the state-of-the-art robust speech recognition system, it is usually beneficial to replace MFCC with MFCC$_{\Delta S}$.

6. REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] B. S. Atal, "Efectiveness of linear prediction characteristics of the speech wave for automatic speaker identication and verication," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[4] S. Furui, "Speaker-independent isolated word recogntion using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 52–59, 1986.

[5] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 1996, pp. 733–736.

[6] Chen Yang, Frank K. Soong, and Tan Lee, "Static and dynamic spectral features: Their noise robustness and optimal weights for ASR," *IEEE Trans. Audio, Speech, and Langugae Processing*, vol. 15, no. 3, pp. 1087–1097, 2007.

[7] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. Eurospeech*, 1991, pp. 413–416.

[8] Jinfu Xu and Gang Wei, "Noise-robust speech recognition based on difference of power spectrum," *Electronics Letters*, vol. 36, no. 14, pp. 1247–1248, 2000.

[9] Heynek Hermansky and Nelson Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[10] Henek hermansky, "The modulation spectrum in the automatic recognition of speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 140–147.

[11] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000, pp. 181–188.