ROBUST SPEECH RECOGNITION USING A VOICED-UNVOICED FEATURE

András Zolnay, Ralf Schlüter and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
52056 Aachen, Germany
{zolnay, schluter, ney}@informatik.rwth-aachen.de

ABSTRACT

In this paper, a voiced-unvoiced measure is used as acoustic feature for continuous speech recognition. The voiced-unvoiced measure was combined with the standard *Mel Frequency Cepstral Coefficients* (MFCC) using linear discriminant analysis (LDA) to choose the most relevant features. Experiments were performed on the *SieTill* (German digit strings recorded over telephone line) and on the *SPINE* (English spontaneous speech under different simulated noisy environments) corpus. The additional voiced-unvoiced measure results in improvements in word error rate (WER) of up to 11% relative to using MFCC alone with the same overall number of parameters in the system.

1. INTRODUCTION

Standard state-of-the-art automatic speech recognition systems use spectral (e.g. Mel Frequency Cepstral Coefficients, MFCC) representation of the acoustic speech signal. Nevertheless these representation techniques are not robust to acoustical variation like background noise, speaker change etc. Word error rate can increase considerably under real life conditions.

A possible way to robust speech recognition could be finding representative features of the speech signal and corresponding robust extraction methods. We tested a voiced-unvoiced measure in the combination of MFCC features with a Hidden Markow Model (HMM) based recognition system. The first related studies go back to rule based speech recognition, where voiced-unvoiced detection was used as one of the different acoustical features, see chapter The Speech Signal in [1]. The method described in [2] utilizes the periodicity of the voiced sound and achieves significant improvement on a Mandarin language task. Several signal analysis based and statistical solutions have already been proposed for voiced-unvoiced detection [3, 4]. A voice onset time based feature is used by [5] in a two-pass recognition system to improve letter recognition accuracy. Recently, articulatory based acoustic modeling techniques apply the voiced-unvoiced feature among other articulatory features [6].

In this paper we report on experiments with a voiced-unvoiced measure for robust speech recognition. Voiced-unvoiced extraction was implemented based on the harmonic product spectrum, see chapter Pitch Detection in [4]. A measure of voicedness is extracted based on the harmonic product spectrum for each time frame. This measure is combined with the standard MFCC features using linear discriminant analysis (LDA). Experiments showed an improvement of up to 11% relative in word error rate due to this single feature.

The rest of the paper is organized as follows. In section 3, the voiced-unvoiced measure will be derived based on the harmonic product spectrum. Two kinds of measures expressing the voicedness of a time frame will be introduced. Experiments will be presented in section 4, followed by the conclusions in section 5.

2. BASELINE SIGNAL ANALYSIS

In this section, the standard short-term power spectrum based signal analysis component of our speech recognition system is described. First we perform a preemphasis of the sampled speech signal. Every 10 ms, a Hamming window is applied to preemphasized 25 ms speech segments. We compute the short-term spectrum by FFT along with zero padding. The number of FFT points is chosen sufficiently high to represent the number of samples in a time frame (e.g. 256 points in case of 8 kHz sampling rate and 25 ms window length). Next, we compute the outputs of mel scale triangular filters, the number of which depends on the sampling rate and varies 15 to 20 in our system. A filter bank is applied to the mel spectrum, in which each filter has a triangular bandpass frequency response with bandwidth and spacing determined by a constant mel frequency interval. For each filter the output is the logarithm of the sum of the weighted spectral magnitudes. Due to

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under the post graduate program “Software für Kommunikationssysteme”.

overlapping filters, filter bank outputs of adjacent filters are correlated. The filter bank outputs are decorrelated by a discrete cosine transform. The optimal number of cepstrum coefficients varies form $M = 12$ to $M = 16$ depending on the number of filters.

Subsequently, a cepstral mean and variance normalization is carried out in order to account for different audio channels. We distinguish two types of normalization: sentence-wise and session-wise. For sentence-wise recorded corpora, normalization is performed on whole sentences. In addition, the zeroth coefficient is shifted so that the maximum value within every sentence is zero (energy normalization). Session-wise recorded corpora consist of recordings containing several sequentially spoken sentences. For these corpora, normalization is carried out with a symmetric sliding window of 2 s without energy normalization. In such way every 10 ms, a vector consisting of normalized cepstrum coefficients is computed.

3. VOICED-UNVOICED FEATURE

Voiced and unvoiced sounds form two complementary classes, thus a feature explicitly expressing the voicedness of a time frame can lead to better discrimination of the phonemes and consequently to better recognition results. Our goal was to find an extraction method which produces a reliable measure of the voicedness of a time frame. For evaluation, we augmented the standard MFCC with this measure.

3.1. Harmonic Product Spectrum

The implemented voiced-unvoiced extraction method is based on the quasi periodic oscillation of the vocal cords. The amplitude spectrum of voiced sounds shows sharp peaks that occur at integer multiples of the fundamental frequency. This fact serves as the basis of the method harmonic product spectrum [4]. The harmonic product spectrum $P(n)$ is the product of R frequency compressed replicas of the amplitude spectrum $|X(e^{jn\Delta\omega})|$, where $\Delta\omega$ is the resolution of the discrete Fourier transform:

$$P(n) = \sqrt[R]{\prod_{r=1}^{R} |X(e^{jn\Delta\omega\ r})|}$$

The motivation for using the product spectrum is that for periodic signals, compressing the frequency scale by integer factors should cause the harmonics to coincide at the fundamental frequency and at its nearby harmonics. Since the amplitude spectrum of a periodic signal is zero between the harmonics, the product of compressed amplitude spectra cancels out all the harmonics falling between two harmonics of the fundamental frequency.

In ideal case the harmonic product spectrum gives high peaks at the fundamental frequency and at its nearby harmonics and it is zero otherwise. Since speech analysis is based on short-time Fourier analysis and even voiced sounds are only quasi periodic, the harmonic product spectrum is not zero between the harmonics of the fundamental frequency and its peaks are not always obvious.

3.2. Measure of Voicedness

The aim of the voiced-unvoiced extraction is to produce a normalized value describing how voiced the current time frame is. We developed two kinds of measures that evaluate the peak structure of the harmonic product spectrum. The measures evaluate the highest point of the harmonic product spectrum. Voiced time frames exhibit a sharp maxima. Sounds with low fundamental frequency can have the maximum point positioned on an obvious peak at the first harmonic frequency. Unvoiced time frames have no clear peak structure and the maxima of the harmonic product spectrum is typically flat. The two kinds of measures capture two different aspects of a peak: height and width.

3.2.1. Height Measure

The height measure v_{height} describes the peak of the harmonic product spectrum by considering only the amplitude of the peak. It is defined as the ratio of the maximum amplitude value at the frequency position n_{max} and the geometric mean of the neighboring amplitudes without the maximum value:

$$v = \frac{P(n_{max})}{\sqrt[2W]{\prod_{n} P(n)}},$$

where n_{max} is the position of the maximum amplitude and the product over n goes through the neighborhood of n_{max} from $n_{max} - W$ to $n_{max} + W$ excluding n_{max}. The size of the neighborhood is chosen to avoid the peak of the first harmonic being included in the average. The minimum pitch and thus the minimum distance between two harmonics is about 80 Hz. W is set such that the size of the neighborhood in both directions is half of the minimum distance between two harmonics, $40Hz/\Delta\omega$.

Typically we have $1 \leq v < 3$. Values $v > 2$ are cut to 2 since they obviously indicate a voiced segment:

$$v_{height} = \min\{2, v\}.$$

3.2.2. Width Measure

A method based on the frequency axis could decouple the measure from the changing loudness and signal to noise ratio of the speech signal. The width measure v_{width} captures the peak of the harmonic product spectrum on the frequency axis. The definition of the width measure is similar to the notion of bandwidth of transfer functions, namely the width w of the peak on the frequency axis at a height of 75% of the maximum amplitude:

$$w = \min\{w\prime : \forall n > w\prime\; P(n_{max} \pm n) < 0.75 P(n_{max})\}.$$

A clear peak gives a value of 1. Peaks wider than an upper bound U are cut to the upper bound. The upper bound is chosen to a value which is surely in the range of the unvoiced time frames, $40Hz/\Delta\omega$. The extracted width values are normalized with the upper bound:

$$v_{width} = \frac{\min\{w, U\}}{U}.$$

3.3. Experimental Setup

The details of generation of the harmonic product spectrum are summarized in this section. Every 10 ms, a Hamming window is applied to the speech signal. The length of the window is in this case larger than for MFCC, 40 ms. To increase the frequency resolution and thus to increase the number of amplitude values between two harmonics, a 2048-point FFT is computed with zero padding. The harmonic product spectrum is composed of the maximum number of compressed amplitude spectra (R). Amplitude spectra can be shrunk up to a width of 400 Hz, since the pitch is lower than 400 Hz. (E.g. by a sampling rate of 8000 Hz $R = 10$.)

4. EXPERIMENTAL RESULTS

Experiments were performed on the small vocabulary corpus *SieTill* and on the large vocabulary corpus *SPINE*. The voiced-unvoiced measure is handled in both cases in the same way. The normalized MFCC feature vectors are augmented with the voiced-unvoiced measure. LDA is applied to choose the most relevant features and to extract the time dependencies. 11 successive augmented vectors of the sliding window $t-5, t-4, ..., t, ..., t+4, t+5$ are adjoined to form a large input vector. The LDA matrix projects this vector onto a smaller dimension subspace by reserving the most relevant classification information. The resulting acoustic vectors are used for recognition.

The baseline experiments apply LDA in the same way. The only difference is in the size of the LDA input vector and thus in the size of the LDA matrix. The resulting feature vector has the same size to ensure comparable recognition results.

The *SieTill* corpus was recorded with 8 kHz sampling rate resulting in 15 mel scale filters and 12 cepstrum coefficients. LDA projects the 11 adjoined feature vectors on a 25-dimensional subspace.

The differences in the *SPINE* corpus are due to the different sampling rate (16 kHz). The wider bandwidth enables 20 mel scale filters and 16 cepstrum coefficients. The 11 adjoined feature vectors are projected by LDA on a 33-dimensional subspace.

4.1. Small Vocabulary Task

The first tests were performed on the *SieTill* corpus [7]. The corpus consists of German continuous digit strings recorded over telephone line: approximately 43k spoken digits in 13k sentences in both training and test set. The number of female and male speakers is balanced.

The baseline recognition system for the *SieTill* corpus is built with whole word HMMs using continuous emission distributions. It is characterized as follows:

- vocabulary of 11 German digits including 'zwo'
- gender-dependent whole-word HMMs with every two subsequent states being identical
- for each gender 214 distinct states plus one for silence
- Gaussian mixture emission distribution and globally pooled diagonal covariance matrix
- 25 acoustic features after applying LDA
- maximum likelihood training using Viterbi approximation

The baseline system has a word error rate of 1.91% which is the best reported so far using MFCC features and maximum likelihood training [7]. In Table 1, the experimental results are summarized for using the additional voiced-unvoiced measure. Experiments were performed with single and with 32 Gaussian densities per mixture. In both cases, a relative improvement in word error rate of 11% is obtained. The tests applying the height and the width based measures did not show any significant differences.

Table 1. Word error rates on the *SieTill* test corpus obtained for MFCC and for MFCC combined with voiced-unvoiced measure (V-U). Experimental results are shown for both measures height and width, as described in section 3.2. #dns gives the average number of densities per mixture.

#dns	acoustic feature	error rates [%]		
		del	ins	WER
1	MFCC	0.49	0.74	3.84
	MFCC + V-U height	0.48	0.41	3.34
	MFCC + V-U width	0.48	0.42	3.33
32	MFCC	0.30	0.52	1.91
	MFCC + V-U height	0.29	0.35	1.70
	MFCC + V-U width	0.29	0.37	1.71

4.2. Large Vocabulary Task

The performance of the voiced-unvoiced measure on a large vocabulary task was tested on the *Speech In Noise*

(SPINE) corpus [8]. The corpus involves human-to-human interaction on a constrained problem solving scenario under six different simulated noisy environments: approximately 12k short sentences from 23 female and 17 male speakers in both training and test set. The baseline recognition system is characterized by:

- recognition vocabulary of 5000 words
- 6-state HMM triphone models with every two subsequent states being identical
- decision tree with 1001 tied states including one silence state
- gender independent Gaussian mixture emission distribution with a total of 80k to 126k densities and globally pooled diagonal covariance matrix
- 33 acoustic features after applying LDA
- maximum likelihood training using Viterbi approximation
- trigram language model with a test set perplexity of 28.5

The baseline system has a word error rate of 30.3 % which has to be compared with word error rates reported by other groups on the same task [9, 8]. These vary from 25.7% to 32.8%. Experimental results are summarized in Table 2. We tested the voiced-unvoiced measure for 80k and 126k total number of densities. A relative improvement in word error rate of 3% is achieved by adding the voiced-unvoiced measure. Differences in the two kinds of measures could not be found under noisy conditions either.

Table 2. Word error rates on the *SPINE* test corpus obtained for MFCC and for MFCC combined with voiced-unvoiced measure (V-U). Experimental results are shown for both measures height and width, as described in section 3.2. #dns gives the total number of densities for 1001 tied states.

#dns [k]	acoustic feature	error rates [%]		
		del	ins	WER
80	MFCC	7.3	4.0	30.3
	MFCC + V-U height	6.9	3.8	29.8
	MFCC + V-U width	7.0	3.9	29.8
126	MFCC	7.0	3.9	30.3
	MFCC + V-U height	6.6	3.7	29.3
	MFCC + V-U width	6.6	3.9	29.4

5. CONCLUSION

In this paper, a voiced-unvoiced measure has been combined with the standard MFCC using LDA. We introduced two kinds of measures (height and width based) of voicedness based on the harmonic product spectrum. Experiments performed on the small vocabulary task *SieTill* achieved an improvement in word error rate of 11% relative compared to the baseline word error rate 1.91%. The large vocabulary tests were performed on the *SPINE* task. The additional voiced-unvoiced measure resulted in an improvement of 3% relative compared to the baseline word error rate of 31.1%.

6. REFERENCES

[1] L. R. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall Signal Processing Series, Englewood Cliffs, NJ, 1997.

[2] L. Gu and K. Rose, "Perceptual Harmonic Cepstral Coefficients for Speech Recognition in Noisy Environment," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001, pp. 125 – 128.

[3] S. Ahmadi and A. S. Spanias, "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333 – 338, May 1999.

[4] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Signal Processing Series, Englewood Cliffs, NJ, 1979.

[5] P. Niyogi and P. Ramesh, "Incorporating Voice Onset Time to Improve Letter Recognition Accuracies," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, pp. 13 – 16.

[6] Y. Gao, R. Bakis, J. Huang and B. Xiang, "Multistage Coarticulatory Model Combining Articulatory, Formant and Cepstral Features," in *Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 25 – 28.

[7] R. Schlüter, W. Macherey, S. Kanthak, H. Ney and L. Welling, "Comparison of Optimization Methods for Discriminative Training Criteria," in *European Conf. on Speech Communication and Technology*, Rhodes, Greece, Sep. 1997, pp. 15 – 18.

[8] NRL SPINE Workshop, "SPINE Evaluation Plan," http://elazar.itd.nrl.navy.mil/spine.

[9] R.C. Rose, Hong Kook Kim and Don Hindle, "Robust Speech Recognition Techniques Applied to a Speech in Noise Task," in *European Conf. on Speech Communication and Technology*, Aalborg, Denmark, Sep. 2001, pp. 2351 – 2354.