Speaker Independent Voiced-Unvoiced Detection Evaluated in Different Speaking Styles

Martin Heckmann[1], *Marco Moebus*[2], *Frank Joublin*[1], *Christian Goerick*[1]

[1]Honda Research Institute Europe GmbH
D-63073 Offenbach am Main, Germany
{martin.heckmann, frank.joublin, christian.goerick}@honda-ri.de

[2]Darmstadt University of Technology, Institute of Telecommunications, Signal Processing Group
D-64283 Darmstadt, Germany
moebus@ieee.org

10.21437/Interspeech.2006-465

Abstract

We propose a new algorithm for voiced/unvoiced classification of speech on a phoneme or sample level. The algorithm is inspired by auditory based approaches and combines two cues. One cue is based on the energy distribution of the signal and the other on the harmonicity. In order to extract the harmonicity of the signal we calculate a histogram of the zero crossings of the filter channels after applying a Gammatone filterbank to the signal. A measure similar to the variance of the zero crossings yields the harmonicity cue. The performance of the algorithm was measured on several minutes of read and spontaneous speech with various speakers. An algorithm proposed by Mustafa et al. [1] served as benchmark. The results show that our algorithm performs significantly better as well on read as on spontaneous speech and seems in particular be better able to to cope with different speaking styles.

Index Terms: speech analysis, voiced/unvoiced detection, speaking style, zero crossings.

1. Introduction

Speech recognition systems that are available today are based on statistical methods (mostly Hidden-Markov-Models) and perform well in situations characterized by stationarity and low-noise conditions. When the distance between the speaker and the microphone gets larger and the noise level increases these systems show dramatic performance drops. Motivated by the fact that humans perform extremely well in such situations we want to incorporate additional features in the recognition process known to be of importance for humans. One such feature is the voicing state or the *Voice Onset Time (VOT)* of a speech segment. Especially measuring the VOT demands for a chronologically very precise determination of the voicing state.

A crucial step in the voicing detection is the selection of the appropriate features. A variety of features has been discussed in the literature [2, 3, 4, 5, 1, 6, 7]. These features rely mainly on two properties of speech segments. Firstly, energy in voiced speech as vowels and consonants, except for plosives, is highly concentrated in the fundamental frequency and its harmonics. Hence the ratio between the energy in the low frequencies to the energy in the high frequencies can be used as an indication on the voicing state of the segment. Secondly, voiced sounds show a harmonic structure whereas unvoiced sounds have a more noise-like distribution of energy across the frequency spectrum. In principle the feature based on the harmonicity of the signal would be sufficient to decide upon the voicing state, but it is very often difficult to asses and prone to errors. Therefore, a combination with an energy based criterion is beneficial.

In the following we will first introduce the features we use and detail how we combine them. Then we will compare our results to a state of the art voiced/unvoiced detection system. This comparison is performed on several minutes of read as well spontaneous speech. After reporting the results of this comparison we close with some concluding remarks.

2. Voicing Detection

For our algorithm we take inspiration from the human auditory system. Consequently, we model some of its properties as for example the preprocessing via a Gammatone filterbank, known to have similar properties as the human cochlea [8]. The filterbank we use has 128 channels ranging from 80 Hz to 5 kHz and follows the implementation given in [9]. In account of the previously mentioned reasons we use a combination of a harmonicity and an energy based feature for the voicing detection.

2.1. The α-ratio

The α-ratio sets the energy in different frequency bands into relation. It can be defined as

$$\alpha = \frac{\int_{f_{\text{high}}}^{\infty} E(f)\,\mathrm{d}f}{\int_{0}^{f_{\text{low}}} E(f)\,\mathrm{d}f}\,, \tag{1}$$

where $f_{\text{high}}, f_{\text{low}}$ denote thresholds for the limits of the considered frequency bands. Usually, the α-ratio is calculated with the whole frequency spectrum under consideration, meaning that $f_{\text{high}} = f_{\text{low}} = f_{\text{thresh}}$, resulting in only one threshold value that is in most cases set around 1 kHz in order to make sure that the main energy of voiced speech is contained in that lower frequency band. As unvoiced speech always has its maximum energy at high frequencies, the α-ratio can be used to differentiate voiced from unvoiced speech. To take into account that vowels sometimes have considerable energy at $1-2$ kHz we use a threshold of $f_{\text{low}} = 1$ kHz for the low frequencies and $f_{\text{high}} = 3$ kHz for the high frequencies. In Fig. 1 the α-ratio is visualized for a test sentence. The phonetic labels are given in the *SAMPA* notation

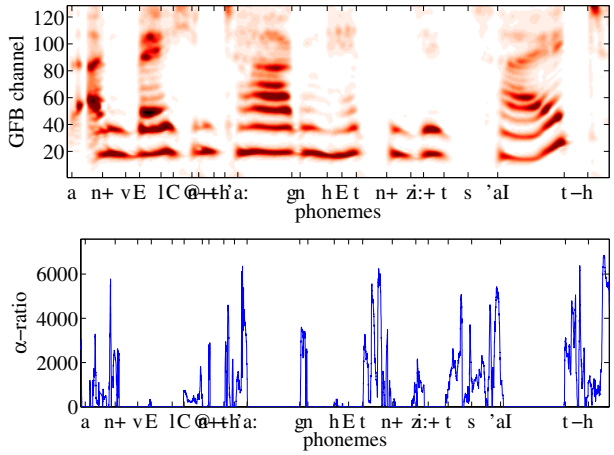

Figure 1: Spectrum (upper plot) and values of α (lower plot) with $f_{\text{high}} = 3\,\text{kHz}$ for the question "An welchen Tagen hätten Sie Zeit?" (an+vElC@n+t-h'a:gnhEtn+zi:+ts'aIt-h).

[10]. Naturally, the values depend heavily on the exact threshold setting, especially the value of f_{high}. The lower f_{high} is set, the more likely it is to include energy in the upper frequency band which stems not from noise-like excitation but from some higher harmonics of a voiced utterance. Likewise when setting f_{high} to a high value unvoiced phonemes with most of the energy concentrated in the mid-range frequency band can not be distinguished from voiced consonants by their α-ratio anymore. We have chosen $f_{\text{high}} = 3\,\text{kHz}$ for our experiments.

2.2. Zero Crossing Distance Histogram

The second, harmonicity based cue evaluates the histogram of the *Zero Crossing Distances (ZCDs)* of the filter signals resulting from the Gammatone filterbank. When signals stem from the same fundamental frequency, they have zero crossings in common. How many zero crossings they share depends directly on their harmonic order relative to the fundamental frequency. For example the first order harmonic shares each second zero crossing with the fundamental. Hence the distance between two zero crossings of the fundamental occurs again as the distance between three zero crossings of the first harmonic and so forth. We want to refer to these distances between multiple zero crossings as higher order zero crossing distances. Not the absolute occurrence of the zero crossings but only their distances can be used due to the frequency and articulation dependent phase delay introduced by the vocal tract. The distance of the fundamental reoccurs in the higher order distances of the harmonics and hence a histogram over all distances shows peaks at the distance value of the fundamental. The energy of the filter signals is represented in the ZCD histogram by weighting the values with the energy of the corresponding channel. This means distance values stemming from a segment with high energy have more weight in the histogram (see [11] for more details). To enhance signal parts with low energy the histogram values are normalized to the maximum at each sample (compare Fig. 2). Unvoiced regions can be identified as regions with energy spread wide across the whole histogram. The pitch contour is clearly visible as it only appears in regions with almost no energy scattering.

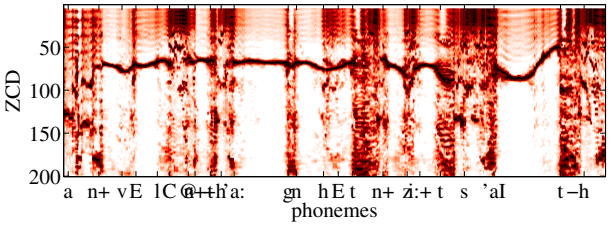

Figure 2: Example of a ZCD histogram for the question "An welchen Tagen hätten Sie Zeit?". The fundamental frequency is depicted as a narrow pitch contour where the signal has high energy.

2.3. Weighted zero crossing distance spread

The harmonicity of the signal segment can now be measured via the distribution in the ZCD histogram. If a sample is part of a periodically excited speech sound, this will be represented by a narrow line in the histogram since almost all energy is contained in the fundamental frequency and its harmonics. In most cases if the harmonic relation of the energy distribution is absent, the sample possesses a distribution where the overall energy is spread across the whole histogram. The contour of this spread represents the distribution function of occurring zero crossing distances and could be interpreted as a *probability density function (pdf)* for the occurrence of these ZCDs. As we don't have access to the true pdf's but only to estimates based on training data we want to refer to these measured distributions $\hat{p}(z)$ as *empirical density functions (edf's)*.

While analyzing the data with respect to their representation in the normalized ZCD histogram, we found that low energy segments in the histogram do not contain information that is specific to the voicing decision. We therefore neglect those segments by filtering the histogram via substituting the energy weight factor $E(z,t)$ for each channel z at time t with

$$E\prime(z,t) = \begin{cases} E(z,t) & \text{if } E(z,t) \geq 0.5 E_{\max}(t), \\ 0 & \text{otherwise} \quad . \end{cases} \tag{2}$$

Following the interpretation of the histogram as empirical distribution functions of ZCDs, this filtering operation is a nonlinear transformation of the *edf*. The output of this filtering operation can be seen in Figure 3.

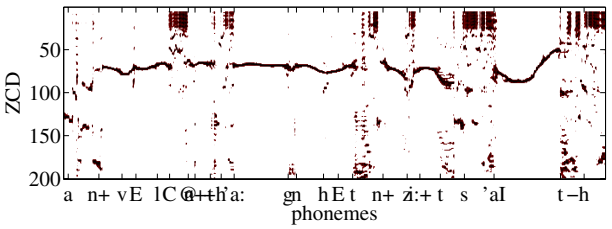

Figure 3: Filtered ZCD histogram

Based on this filter output, one can calculate a measure of concentration. The feature we finally extract out of the filtered histogram is defined as

$$\sigma^2_{\mathrm{CoG}} = \frac{\int_{ZCD_{\min}}^{ZCD_{\max}} \hat{p}(z)E'(z)(z-\mu)^2\,\mathrm{d}z}{\int_{ZCD_{\min}}^{ZCD_{\max}} \hat{p}(z)E'(z)\,\mathrm{d}z} \quad (3)$$

where μ is the first moment of the transformed *edf* such that

$$\mu = \frac{\int_{ZCD_{\min}}^{ZCD_{\max}} \hat{p}(z)E'(z)z\,\mathrm{d}z}{\int_{ZCD_{\min}}^{ZCD_{\max}} \hat{p}(z)E'(z)\,\mathrm{d}z}\,. \quad (4)$$

In this respect σ^2_{ZCD} is very similar to the variance of the ZCDs. In Figure 4, the feature plot of σ^2_{ZCD} is shown. The difference between voiced and unvoiced speech is clearly mapped into a mostly disjunctive feature domain. Voiced speech that is excited periodically results in very low values of σ^2_{ZCD} whereas unvoiced speech shows feature values up to $\sigma^2_{\mathrm{ZCD}} = 10000$.

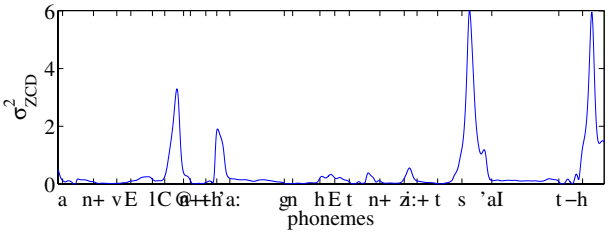

Figure 4: Example of σ^2_{ZCD} values for a speech signal .

For unvoiced speech which is not coarticulated and therefore clearly does not contain a fundamental frequency, σ^2_{ZCD} can be small although there is no concentration of energy in single channels. This may happen if there is only activity in the very highest frequencies of the spectrum which results in a very skew distribution in the histogram. To avoid errors in these cases, we introduce a small error-correction that checks for the largest peak width $W_{\max}(t)$ at the present sample. We define the widest peak as the segment covering the greatest number of channels in the histogram after it has been filtered according to Eq. (2).

Since a small σ^2_{ZCD} has to correspond also to a narrow pitch contour in voiced speech, this is not satisfied in the above described cases. We therefore override the corresponding σ^2_{ZCD} to make sure the sample is not regarded as voiced. The threshold for this correction was found empirically as a maximum peak width of 10 channels, such that the error correction can be expressed as

$$\sigma'^2_{\mathrm{ZCD}}(t) = \begin{cases} \sigma^2_{\mathrm{ZCD}}(t), & \text{if } W_{\max}(t) > 10 \\ \xi & \text{otherwise} \end{cases} \quad (5)$$

with ξ set to an artificial high value, e.g. $\xi = 1000$.

3. Feature Integration and Decision System

For the integration of the two cues we formulate the problem as a multidimensional hypothesis test. The dimensions along which the test operates are the two cues $\mathbf{x}(\alpha(t), \sigma^2_{\mathrm{ZCD}})$ and the hypotheses H_i are voiced V or unvoiced UV. In order to do so we have to estimate the likelihood $p(\mathbf{x}|H_i)$. We use 34 seconds of spontaneous speech using the phonetically labeled *Kiel Corpus of Spontaneous Speech* [12] for this estimation. Likelihoods were estimated for a male and a female speaker but no significant difference was found. The decision upon voicing was done based on a rectangular decision region in the two dimensional feature domain Θ

$$R_V = \{(\alpha, \sigma'^2) \in \Theta \quad | \quad \alpha \leq 0.5 \wedge \sigma'^2_{\mathrm{ZCD}} \leq 100\}, \quad (6)$$

$$R_{UV} = \{(\alpha, \sigma'^2) \notin R_V\} \quad (7)$$

The decision is made for individual samples and afterwards, a median filtering is applied with a sliding window of 10 ms.

4. Results

The detection performance of the developed system was assessed via a large set of data from the *Kiel Corpus of Read and Spontaneous Speech*. To measure the influence of speaking style and speaker variability on detection performance, we calculated statistics for read and spontaneous speech separately.

For read speech, we used a set of 20 sentences that were spoken by 6 different speakers (3 female, 3 male), resulting in a total of 389 seconds of analyzed speech. Spontaneous speech was analyzed by a set of 76 utterances from 4 speakers (2 female, 2 male), with a total of 287 seconds. The utterances were recorded while the speakers had to accomplish an appointment-making task. The files are phonetically transcribed and time aligned.

Since no explicit voicing label is available, we obtain the voicing information by mapping the phonemes to a voicing status according to their phonological classification. For plosives the label was split up into two parts, since they are produced with an initial closure of the glottis, followed by a noise burst. For technical reasons the closure phase was assumed to be voiceless for all phonemes, thus not taking into account that voiced plosives show voicing in the closure phase. Therefore, the first part of the phoneme, corresponding to the closure phase, was labeled as silence and not included in the evaluation. The voicing of the second part of the plosive, the release and (for voiceless stops) aspiration phase, was then labeled according to phonology.

This allows us to automatically process and evaluate a larger amount of data, but also has the drawback that this does not take phoneme dependent VOTs and coarticulation into account. Since the voicing of speech will not always change with phoneme borders but will also depend on context, the mapping of phonemes to voicing is therefore not optimal.

This is especially true for spontaneous speech where pronunciation is less careful. To include these contextual effects in the label, we also labeled a set of spontaneous speech *manually* by the inspection of waveforms and spectra of the signals. This increases the precision and therefore the quality of the reference labels. The set of manually labeled speech contains 34 seconds of spontaneous speech, recorded from one female and one male speaker who have not yet been included in the data set.

We compared the results we obtained to an algorithm developed by Mustafa et al. [1].

4.1. Algorithm of Mustafa et al.

This algorithm is based on two features: the α-ratio and the autocorrelation. The α-ratio α_{Std} used by Mustafa et al. uses only one cut-off frequency. Additionally a gender detector based on a pitch tracking is implemented. Via the use of the gender detector the parameters of the features, especially the cut-off frequency used in the calculation of the α-ratio, are adapted. The use of the auto-

correlation as a feature is motivated by the fact that voiced sounds have a systematically higher autocorrelation than unvoiced sounds. An additional hysteresis block avoids fast changes of the voicing state due to measurement errors. Each of the two features decides about the voicing independently. To obtain a final decision, these decisions are then combined by the application of some rules, i.e. the segment is regarded as voiced if the hysteresis block detects voicing and the autocorrelation is above a minimal value.

4.2. Comparison

During the comparison we used for both algorithms the speech data detailed above. The decision regions used in our algorithm were set as given in the previous section and not changed during the different tests. Since the algorithm proposed by Mustafa et. al. sometimes fails to detect speech onsets, we do consider only those speech signals in the statistic, where the tracking caused no errors, i.e. it was able to detect at least 50% of the voiced samples. The statistics in the following therefore reflect only differences in the detection performance due to feature selection and combination, not differences that occur due to tracking errors.

4.3. Read Speech

The results for the read speech part are shown in Tab. 1. The values

	Total	Vowels only	Consonants only
Our algorithm	82.4 %	86.4 %	79.5 %
Mustafa et al.	71.5 %	81.3 %	64.4 %

Table 1: Correct classification rates for the read speech part, measured on phonological voicing labels.

are given in percent of samples labeled as speech correctly classified. As can be seen our algorithm clearly performs better overall and especially for consonants.

4.4. Spontaneous Speech

For spontaneous speech we performed two tests. One where we used the phonological voicing labels and one where we adjusted the labels by hand. In Tab. 2 the detection rates for the phonologically labeled part are given. In this case again the differences be-

	Total	Vowels only	Consonants only
Our algorithm	80.7 %	83.3 %	79.0 %
Mustafa et al.	60.8 %	61.7 %	60.3 %

Table 2: Correct classification rates for the spontaneous speech part, measured on phonological voicing labels.

tween our algorithm and that of Mustafa et al. are even larger. Due to the less precise articulation in spontaneous speech the detection rates, especially for vowels, degrade. Finally we give the results for the spontaneous speech part with manually adapted voicing labels in Tab. 3. In this test the effects of the imprecise labeling are

	Total	Vowels only	Consonants only
Our algorithm	86.3 %	89.2 %	84.0 %
Mustafa et al.	68.2 %	71.7 %	65.4 %

Table 3: Correct classification rates for the spontaneous speech part, measured on manually obtained labels.

reduced and the performance of the algorithm can be better assessed. For the manual set labels our algorithm performs best and is again much better than that of Mustafa et al.

5. Conclusion

We developed an algorithm for voiced/unvoiced decision which operates on the sample level. As features we use the energy distribution and a measure based on the entropy of the zero crossing distances of the signal after the application of a Gammatone filterbank. The comparison to a state of the art algorithm by Mustafa et al. showed that our algorithm performs significantly better at less computational costs (the calcuation of the autocorrelation and the gender detector in the algorithm of Mustafa et al. are computationally quite demanding). The tests were performed for two different databases, one with read speech and one with spontaneous speech. From both databases we used a large set of utterances. We could show that the voicing detection in vowels works better for read speech than for spontaneous speech. This is most likely due to coarticulation effects, especially the vowel length reduction in spontaneous speech. We made an additional test on the spontaneous speech corpus where the labels were adjusted by hand. The significant differences to the test on the purely phonetic labels illustrates their drawbacks for such a task. They especially do not capture coarticulation effects, which yield deviations between the phonetic identity of a segment and its actual voicing state. On this spontaneous speech set with manually set labels our algorithm obtained the best results with detection rates well above 80%.

6. References

[1] Kamran Mustafa and Ian C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Speech and Audio Proc.*, 2006, accepted.

[2] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 3, pp. 333 – 338, May 1999.

[3] W. A. Ainsworth, "Some approaches to automatic speech recognition," in *The Handbook of Phonetic Sciences*, pp. 721–744. Blackwell, 1999.

[4] A. M. A. Ali, J. van der Spiegel, and P. Mueller, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Proc. (ICASSP)*, Seattle, WA, 2001, pp. 961 – 964.

[5] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 24, no. 3, pp. 201 – 212, 1976.

[6] L. Siegel and A. Bessey, "Voiced/unvoiced/mixed excitation classification of speech," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Proc. (ICASSP)*, Paris, 1982, pp. 451 – 460.

[7] J. A. Marks, "Real time speech classification and pitch detection," in *Proc. of the Southern African Conf. on Communications and Signal Proc. (COMSIG)*, Pretoria, South Africa, 1988, pp. 1–6.

[8] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C.Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, Proceedings of the 9h International Symposium on Hearing*, Y Cazals, L. Demany, and K. Horner, Eds., Pergamon, Oxford, 1992, pp. 429–446.

[9] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.

[10] J. Wells, "SAMPA computer readable phonetic alphabet," www.phon.ucl.ac.uk/home/sampa.

[11] Martin Heckmann, Frank Joublin, and Edgar Körner, "Sound source separation for a robot based on pitch," in *Proc. Int. Conf. on Intelligent Robots & Systems (IROS)*, Edmonton, Canada, 2005, pp. 203–208.

[12] Klaus J. Kohler, "Labelled data bank of spoken standard german - the kiel corpus of read/spontaneous speech," in *Proc. Int. Conf. Spoken Language Proc. (ICSLP)*, Philadelphia, PA, 1996, pp. 1938–1941.