A continuous voicing parameter in the frequency domain

L. Koenig$^{(1,2,3)}$, *R. André-Obrecht*$^{(1)}$, *C. Mailhes*$^{(2)}$ *and S. Fabre*$^{(3)}$

$^{(1)}$ University of Toulouse, IRIT, 118 Route de Narbonne, F-31062 Toulouse, France
$^{(2)}$ University of Toulouse, IRIT, 2 rue Camichel, F-31071 Toulouse, France
$^{(3)}$ Freescale Semiconductor, 134 Avenue du Général Eisenhower, F-31023 Toulouse, France
{lionel.koenig, serge.fabre}@freescale.com, corinne.mailhes@enseeiht.fr, obrecht@irit.fr

Abstract

In automatic speech analysis, voicing articulatory is often defined as a binary decision: voiced or unvoiced. Linguists agree that this articulatory should be continuous. In this paper, we present a new approach to compute a continuous voicing indicator of a speech frame. This voicing percentage is then evaluated in both a segmentation process and a speech recognition task. Promising results show that this continuous voicing percentage may be used as a reliable voicing indicator.

1. Introduction

Reliable speech voicing indicator is of great importance in speech recognition [1], lost speech frame estimation [2] or speech synthesis. The aim of this indicator is to determine if the speech produced involves the vibration of the vocal chords, the sound is said voiced, or not (unvoiced). Many approaches address this topic using pitch detectors [3, 4], while other studies are based on classification [5, 6].

Using pitch estimation as a voicing information or a binary classification leads to some problems to introduce this additional information in continuous models such as classical continuous Hidden Markov Models. To get round this difficulty, it is often simpler to consider two different models: one for voiced frames and another one for unvoiced frames [2]. Transitions between voiced and unvoiced frames are handle with rules.

Specialists in linguistics agree that the voice articulatory feature is not binary as many articulatory features; various degrees of voicing are observed. So it will be interesting to have an efficient continuous representation of the voicing; such a parameter may be used easily in speech processing for recognition or prediction.

Only few continuous voicing features exist: among the most common used, one can cite the periodicity[1], the jitter[1] or the minimum of cumulative mean normalized difference[4].

In the present paper, we present a study about a continuous voicing indicator based on a classical spectral representation of the speech. Estimation is provided using a non linear filtering of the Power Spectral Density (PSD). This feature, named the voicing percentage, can be either used as a stand-alone tool for segmenting speech in voiced / unvoiced parts, but it will be certainly more interesting if it may be introduced in a speech processing system: it may be used to adaptively weight the components of a Gaussian Mixture Model for speech recognition or speech reconstruction, or to define continuous articulatory movements in an inverse acoustic-articulatory problem. So to assess this new feature in such various situations, we propose two experimental protocols. One is dedicated to the binary situation where a threshold of the voicing percentage leads to a voiced/unvoiced segmentation; the other one consists of introducing the voicing percentage in an acoustic-phonetic decoder.

Consequently, the next section presents the proposed voicing percentage while throughout the third and fourth sections are described the two experimental protocols and their results.

2. Voicing feature extraction

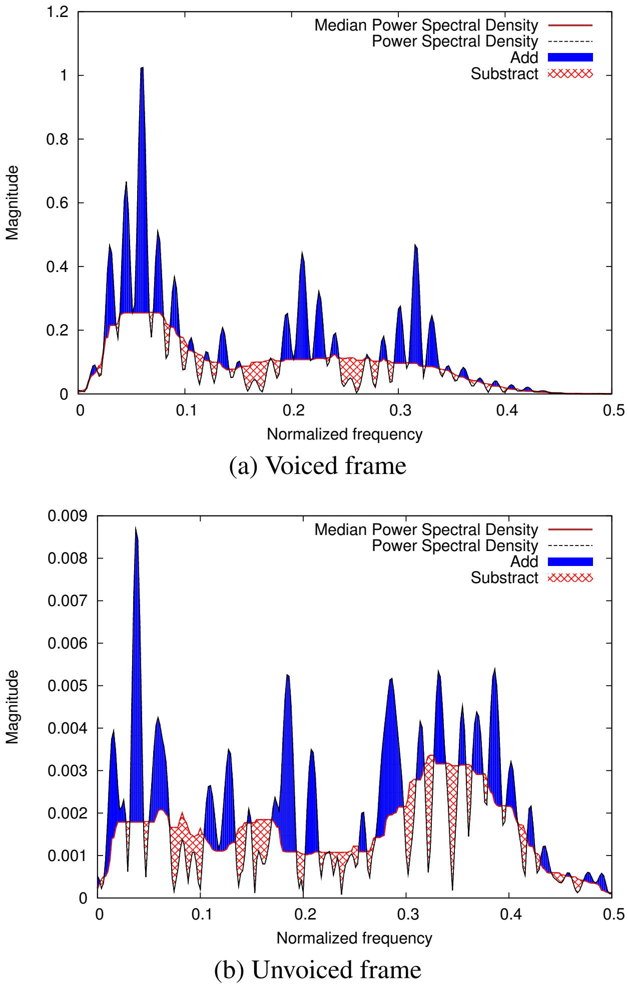
The voice feature, called the voicing percentage and noted $v_{\%}$, is defined as the ratio between the "voicing power" and the overall power of the analyzed speech frame. Voicing power is estimated as the power of the signal frame minus the power of its noise part. To evaluate the spectral part of the noise in the PSD estimate, we propose to extract the basis line of the PSD using a one dimension median filter applied directly on the PSD. Integral of this quantity leads to an estimation of the noise power.

The voicing percentage is thus defined as

$$v_{\%} = \frac{\int_0^{0.5} \left(S(\tilde{f}) - \mathit{median}[S](\tilde{f}) \right) \mathrm{d}\tilde{f}}{\int_0^{0.5} S(\tilde{f}) \mathrm{d}\tilde{f}} \tag{1}$$

where $\tilde{f}$ is the normalized frequency and $\mathit{median}[S](\tilde{f})$ denotes the output of the median filter applied on the PSD.

Figure 1 illustrates this voicing percentage computation on a voiced frame (a) and an unvoiced one (b).



(a) Voiced frame

(b) Unvoiced frame

Figure 1: Examples of voicing percentage estimation.

On this figure, the solid line represents the output of the median filter, while the integral of the solid part of the PSD minus the hatching part corresponds to the numerator of (1).

One effect of the median filter on the PSD is to remove the DC component. To avoid bias introduced by non-zero-mean frames, we remove the offset before the spectral analysis. In our application context, speech frames are classically composed of only $20ms$ of signal. Therefore, a simple periodogram is used as a PSD estimator, including a windowing (Blackman window) of the signal in order to have deeper harmonic gaps in voiced frames. Thus, the effect of the median filter is more efficient.

Figure 2 sums up the voicing percentage estimation algorithm.

In practice, we rapidly remark this indicator is of no sense when the speech frame is silence. In this case, theoretically, the algorithm leads to a non defined operation: $0/0$; in pratice, it leads to ratio near to $0/0$, due to the power of the frame, close to zero. To improve the efficiency of this indicator, it is coupled with a simple voice activity detection based on an energy threshold.

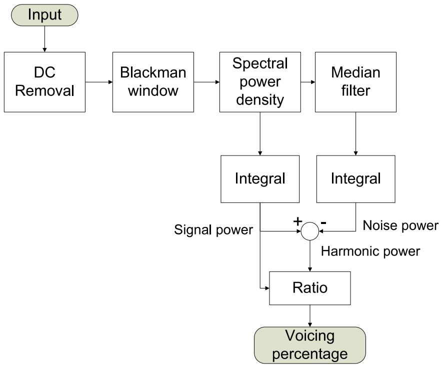


Figure 2: Voicing percentage estimation algorithm.

3. Voiced/unvoiced Segmentation

Including the voicing percentage in a segmentation process implies to define a threshold. This will be discussed in paragraph 3.1. Moreover, a smoothing to eliminate short aberrant decisions is often necessary. To avoid this post processing, we propose, in the paragraph 3.2, an alternative approach where an *a priori* segmentation of the signal is performed to locate stable segments and each segment is *labeled a posteriori* as voiced/unvoiced by using the voicing percentage feature.

A first evaluation was made by comparing the segmentation produced by thresholding the voicing percentage, and the one obtained by mapping the hand-made phonetic labels from OGI to voiced or unvoiced labels. Then a full evaluation is performed using a pitch based automatic segmentation as reference in the paragraph 3.3.

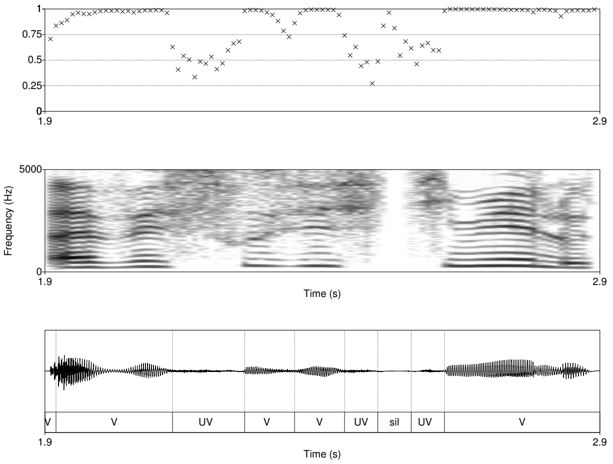
For the assessment, we have used two well-known databases:

- the phonetic segmented part of the OGI Multilingual Telephonic Speech corpus (six languages: English, German, Spanish, Mandarin, Japanese, Hindi) [7]. The sample rate is 8kHZ and the quality is a telephonic one.

- the French corpus BREF80 [8] (table 3 sums up the BREF80 corpus). The corpus has been automatically labeled in phonetic codes [9]. The sample rate is 16 kHZ and the quality is quite good.

3.1. Threshold on voicing percentage

The proposed voicing percentage can be used as a standalone tool for segmenting speech. The segmentation is obtained by thresholding this voicing percentage for voiced / unvoiced decision and thresholding the energy for voice activity detection. Each decision is taken on a frame, no smoothing is performed.



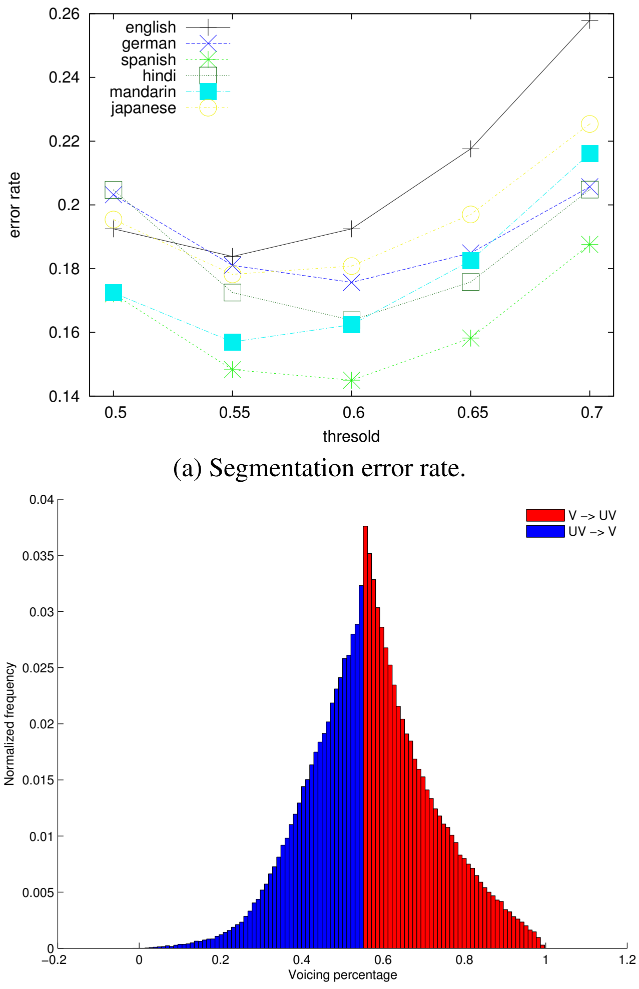
From bottom to top : waveform and the resulting segmentation and labeling, spectrogram and voicing percentage $v_{\%}$

Figure 3: Frame based segmentation.

Figure 3 illustrates the resulting segmentation, based on this voicing percentage threshold. In a first step, this segmentation is compared to the one obtained from the phonetic labeling of the OGI corpus, from which a mapping has been made to voiced / unvoiced and silence. To quantify the assessment, an error rate has been defined as the ratio between time where both segmentations mismatch and overall segmentation time [10]. Figure 4(a) shows the evolution of this error rate as a function of the threshold value. An optimal value of the threshold is around 0.55 and 0.6. Moreover, we are interested in analyzing the statistics of the voicing percentage during wrong segmentation time in order to better understand the reason of this mismatching.

Figure 4(b) presents an histogram of voicing percentage values corresponding to mismatching between both segmentations. The shape of this histogram confirms that wrong segmentations are due to the threshold level: when a wrong decision is taken, values are close to the threshold level.

However, this first segmentation comparison has been made using the OGI labeling mapping as a reference segmentation. It is interesting also to compare the proposed segmentation with an automatic one, non depending on any corpus labeling. Therefore, we propose to use as a second segmentation reference, the results obtained by using a pitch estimator, such as the YIN one [4]. Results in terms of the previously defined error rate between the proposed segmentation and the two reference segmentations are given in table 1. Note that results are quite comparable using any reference segmentation. Therefore, in what follows, the segmentation based on the YIN pitch estimator will be preferred as the reference segmentation, since this is an automatic one which can be used on any corpus.



(a) Segmentation error rate.

(b) Wrong segmentation value histogram.

Figure 4: Voicing percentage evaluation.

3.2. A posteriori voicing percentage segmentation

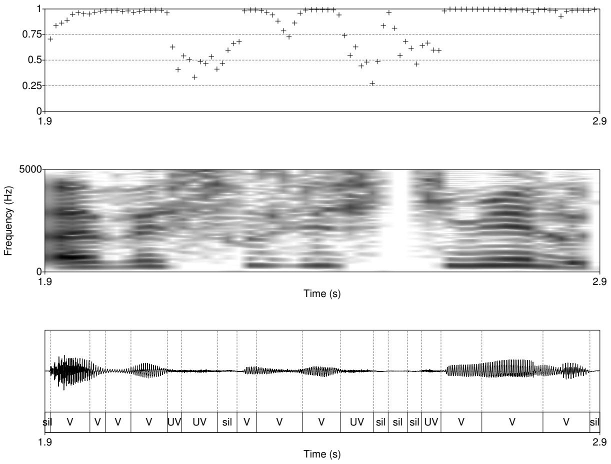
Another way of using the proposed voicing percentage in speech segmentation is presented hereafter. An a priori automatic segmentation is performed on the speech signal. Infra phonetic segments are produced, they correspond to stable units and transient ones [11]. As in the previous system, silence/speech decision is based on a threshold on the power of the segment. For each speech segment, the decision voiced/unvoiced is taken according to the voicing percentage of the $20ms$ middle part of the segment (figure 5).

Thus, the voicing percentage is no more a stand-alone tool for segmentation but, rather, a way to improve an a

Langage	**References**	
	OGI labeling	YIN
English	18.38	19.34
German	18.10	17.83
Spanish	14.83	16.46
Hindi	17.25	15.59
Mandarin	15.70	16.19
Japanese	17.82	16.14

Table 1: Frame based segmentation error rate in %.

priori segmentation, by adding a labeling on each segmented part.



From bottom to top : segmented waveform and the resulting labeling, spectrogram and voicing percentage $v_{\%}$

Figure 5: Voicing percentage labeling with a priori segmentation.

Note that the segmentation seems more accurate, allowing to detect some silence part more precisely. Comparison with the chosen reference segmentation is conducted in the next paragraph.

3.3. Comparison with a pitch based automatic segmentation

Thus, the proposed voicing percentage can be used either as a stand-alone tool for speech segmentation, either as a complementary tool to improve some a priori segmentation. Both using are compared with the segmentation produced by an automatic pitch estimator. In this reference segmentation, the YIN [4] algorithm is used for extracting the fundamental period. The voiced/unvoiced detection and the pitch extraction are completely tied, the pitch is computing over a frame if and only if the frame is estimated voiced. The decision is taken over each frame

Corpus	Language	Stand-alone	A posteriori
	English	19.34	16.57
	German	17.83	15.47
OGI	Spanish	16.46	14.45
	Hindi	15.59	14.35
MLTS	Mandarin	16.19	14.22
	Japanese	16.14	15.04
	All	17.00	15.10
BREF80	French	10.74	13.77

Ratio in % between wrong segmented time and overall segmented time using YIN automatic segmentation as reference

Table 2: Segmentation results

as in paragraph 3.1. Results are given in table 2, in terms of the error rate defined in paragraph 3.1. Both using of the voicing percentage are evaluated and compared: 'stand-alone' stands for using the voicing percentage as a stand-alone tool for speech segmentation, as presented in paragraph 3.1, while 'a posteriori' corresponds to an a posteriori using of the voicing percentage for improving some pre-segmentation, as detailed in paragraph 3.2. The binary implementation of the voicing percentage succeeds in reaching quite good performance. These results are very promising which lead us to investigate the interest of such an indicator as a fully continuous parameter.

4. Voicing percentage and acoustic-phonetic speech recognition

Before using the voicing percentage in speech prediction or inverse problem, it appears necessary to assess its influence in a classical speech recognition system, more precisely in an acoustic-phonetic decoder. Most of the acoustic-phonetic decoders are based on Hidden Markov Models (HMM); the introduction of non homogeneous features in the observation vectors, as articulatory features, doesn't guarantee an improvement in terms of performance. We have elaborated an experiment to quantify the consequence of adding the voicing percentage in the feature vector, in such a decoder.

4.1. Baseline decoder

The acoustic-phonetic decoder is based on the classical HMM framework. As we use the corpus BREF80, a French corpus, 35 phones are defined as in [12]. An HMM with three states is proposed for each phone and the observation pdf is assumed to be a Gaussian Mixture Model with thirty-two components. To train the 35 context-independent phone models on the French corpus BREF80, the same phonetic labels as in the paragraph 3 are used. The repartition between the train corpus and the test corpus is detailed in the table 3. Note that none phonetic grammar is considered.

Part	Subpart	Length
Train	male	4:33:12
	female	5:41:45
	total	10:14:57
Test	male	0:27:46
	female	0:29:58
	total	0:57:44

Table 3: BREF80 Corpus

The observation vector is composed of 12 linear predictive cepstral coefficients, energy, deltas; a cepstral subtraction is applied.

The development of this acoustic phonetic decoder has been made using the HTK toolbox.

Performances of the baseline phonetic decoder (see table 4) are similar to the state-of-art ones[12], when no grammar is introduced. The accuracy value is about 59,9% while the phone correct rate (PCR) is around 67,9%.

4.2. Impact of the voicing percentage

To assess the impact of the voicing percentage feature in the recognition process, it has been added to the vector observation. Thus, the observation vector includes linear predictive cepstral coefficients, energy, deltas, cepstral subtraction *and voicing percentage*.

The learning process was similar to the baseline phonetic decoder.

The introduction of the voicing percentage into the feature vector leads to a small gain in the phonetic recognition (table 4). The accuracy value is about 60.4% while the phone correct rate is around 68.4%. This result is very interesting because it shows the compatibility between non homogeneous observations and offers new perspectives to exploit this property.

Model	Accuracy	PCR
Baseline decoder LPCESTRA_E_D_Z	59.92%	67.92%
Proposed decoder LPCESTRA_E_D_Z + $V_{\%}$	60.39%	68.42%

Table 4: Phonetic speech recognition rates

5. Conclusion

In this paper, we have presented a new continuous voicing estimator which can be used either as a voiced/unvoiced segmentation tool or as a new feature in speech processing based on Hidden Markov Models.

Promising results shown by this continuous voicing percentage may be used as a reliable voicing indicator. Future work will explore its potentiality to adapt the GMM weights in a speech recognition system and to guide a speech prediction process in the case of inverse acoustic-articulatory and packet loss concealment problems.

6. References

[1] D. L. Thomson and R. Chengalvarayan, "Use of voicing features in HMM-based speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 197 – 211, 2002.

[2] C. Rodbro, M. Murthi, S. Andersen, and S. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, vol. 14, no. 5, pp. 1609–1623, 2006.

[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] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, April 2002.

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

[6] I. M. Khademul, H. Keikichi, and M. Nobuaki, "Robust voiced/unvoiced speech classification using empirical mode decomposition and periodic correlation model," *Interspeech conference proceeding*, vol. 1, p. 1, 2008.

[7] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multilanguage telephone speech corpus," in *International Conference on Speech and Language Processing*, oct 1992, pp. 895–898.

[8] L. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *Proceedings of the European Conference on Speech Technology, EuroSpeech*, Genoa, Sep. 1991, pp. 505–508.

[9] O. Le Blouch and P. Collen, "Automatic syllable-based phoneme recognition using ESTER corpus,"

in *ISCGAV'07: Proceedings of the 7th WSEAS International Conference on Signal Processing, Computational Geometry & Artificial Vision*, 2007, pp. 81–85.

[10] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2–3, pp. 225–254, Jun. 2000.

[11] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 1, pp. 29–40, Jan. 1988.

[12] J.-L. Gauvain and L. F. Lamel, "Speaker-independent phone recognition using BREF," in *Proceedings of DARPA Speech and Natural Language Workshop*, Feb. 1992.