Pitch Synchronous Spectral Analysis for a Pitch Dependent Recognition of Voiced Phonemes - PISAR

Hans-Günter Hirsch
Institute for Pattern Recognition, Niederrhein University of Applied Sciences, Krefeld, Germany
hans-guenter.hirsch@hs-niederrhein.de

Abstract

Humans use the pitch of their conversational partner as an important feature for improving the communication and the understanding especially in noisy situations. This knowledge is taken to investigate the idea of a pitch synchronous spectral analysis and a pitch dependent recognition of voiced speech segments. A first approach is presented for realizing this pitch dependent processing. Its applicability is shown for recognizing the voiced segments of the Timit database.

Index Terms: pitch synchronous spectral analysis, pitch dependent recognition

1. Introduction

The recognition of noisy speech can be considerably improved by applying a speech analysis scheme to extract robust acoustic features, e.g. [1], or adapting the reference patterns, usually Hidden Markov Models (HMMs), to the noisy scenario, e.g. [2],[3]. The author developed exemplary solutions for both approaches during the past decades [4], [5]. We observe a certain saturation of the achievable improvement independent of the applied technique.

In almost all noise scenarios the performance of automatic recognition systems is worse than the one of humans communicating under similar noisy conditions. Observing the communication between humans in a fairly noisy situation, the listener perceives and understands with a high accuracy these speech segments that are characterized by a high sound level. The term signal to noise ratio (SNR) has been introduced to include the relation to the noise level. Such segments with a high SNR have also been characterized as "islands of reliability" [6]. Regarding the remaining speech fragments, the listener is only able to extract some more or less unreliable information from certain parts of these fragments. The mentioned observation in human communication was the initial reason to just recently start a project on developing and investigating an alternative recognition scheme. This scheme is based on the detection of speech segments with a high SNR. The intention is similar than the one of these approaches that are known under the term "missing data theory" [7]. The new aspect is the development and application of a different recognition strategy. We intend to start the calculation of the probabilities at the detected voiced segments. Starting at these segments, we want to separately calculate accumulated probabilities backward and forward in time. Furthermore, like with the missing data approaches we want to ignore or take into account only with a lower priority these fragments that have a fairly low SNR. An open issue is the combination of all the probabilities calculated for the fragments between the detected voiced segments to get a total probability and thus a recognition result for a whole utterance. As first step towards the mentioned alternative recognition scheme, we are developing an algorithm to detect voiced segments with a high SNR within noisy speech.

Furthermore, we started thinking about an alternative analysis technique to extract acoustic features for the robust recognition of the detected voiced segments. Usually, parameters like the Mel frequency cepstral coefficients (MFCCs) and the logarithmic short-term energy (logE) are used as acoustic features. The MFCCs contain the information about the transfer function of the vocal tract. Analyzing the perception of voiced sounds, humans extract and use the pitch information to roughly classify the speaker in certain pitch categories. Furthermore, they probably adapt their analysis and recognition processing to this pitch. Approaches known under the keyword "vocal tract length normalization" try to cover these aspects to a certain extent, e.g. [8].

Based on the mentioned knowledge about the human perception of voiced sounds we are developing an analysis scheme that contains the detection of successive pitch periods and the spectral analysis of these pitch periods by applying a DFT (Discrete Fourier Transform). A lot of approaches exist for detecting the periods of a voiced phoneme, e.g. [9], [10]. The analyzed sequence of pitch spectra is taken to perform a pitch dependent recognition of each voiced segment. The recognition task is the classification of each segment as one of 15 classes containing voiced phonemes. Therefore, we split the pitch frequency range from about 70 to 330 Hz in 15 regions. For each pitch region, we estimate and provide individual reference patterns. Dependent on the analyzed pitch value, the recognition is performed by calculating the probabilities to generate the observed sequence of pitch spectra with the corresponding HMM set for this pitch region.

In the next section, we present details about the pitch synchronous spectral analysis scheme. Then we describe the pitch dependent recognition approach. First recognition results are presented by applying the new techniques on data of the Timit data base [11]. In general, we took versions of the Timit speech signals downsampled at a rate of 8 kHz. We compare the results to the case of using "standard" acoustic features like MFCCs and the frame energy.

2. Pitch Synchronous Spectral Analysis

As mentioned before, we are working on a processing scheme to detect voiced segments with a high SNR. We achieve already fairly good detection results with our current version. But to be independent of the accuracy of the detection algorithm, we assume an ideally working detection for the investigations within this paper. Therefore, we use the available labeling information of the Timit data base. We take the phoneme set of the CMU dictionary consisting of 39 phonemes [12]. We applied the rules as defined by [13] to reduce and map the set of 52 phonemes of the Timit data base to the 39 phonemes of the CMU dictionary. These rules are taken to create the labeling information for the CMU phoneme set based on the available labeling of Timit.

Table 1 contains the set of 15 phonemes that we consider and define as voiced phonemes.

10.21437/Interspeech.2013-430

aa	ae	ah	ao	aw	ay	eh	er
ey	ih	iy	ow	oy	uh	uw	

Table 1. *Set of 15 voiced phonemes.*

We applied our robust Mel cepstral analysis scheme [4] on the 4620 Timit utterances designated for training a recognition system. There exist more than 55000 segments within the 4620 utterances containing one the 15 voiced phonemes. Besides using the MFCCs and the frame energy for the recognition experiments later on, our intention is the use of the frame energy to define the starting point of our pitch synchronous analysis. We look at this 25 ms speech frame within each voiced phoneme segment that has the highest frame energy logE. Taking the sample index n_{mid} in the middle of this frame we calculate a first rough estimate $kpitch_{est}$ of the pitch length by a cepstral analysis of 256 samples around the middle index. We search for the maximum cepstral coefficient at a pitch length between 25 and 113 samples. This pitch length corresponds to pitch frequencies between about 70 and 320 Hz at the sampling rate of 8 kHz.

With this first estimate $kpitch_{est}$ we apply a correlation based processing to detect individual pitch periods before and after the index n_{mid}. To detect preceding pitch periods we apply the loop processing shown in figure 1.

$$
\begin{aligned}
&kp = kpitch_{est}\\
&n_{start} = n_{mid} - kp/2\\
&do \quad \{\\
&\qquad [ccfmx, mxind] =\\
&\qquad \underset{rg=-5,\ldots,5}{MAX}\left\{\sum_{k=0}^{kp+rg-1} s(n_{start}+k)\cdot s(n_{start}-(kp+rg)+k)\right\}\\
&\qquad kp = kp + rg(mxind)\\
&\qquad n_{start} = n_{start} - kp\\
&\qquad save\ sample\ indices\ of\ \ pitch\ in\ \ pind()\\
&\qquad \} while(ccfmx < thres1)
\end{aligned}
$$

Figure 1: *Correlation based detection of pitch periods.*

The similarity between two pitch periods is estimated by calculating the correlation product with the corresponding speech samples s(n). We allow a slight variation of the estimated pitch length in the range of -5 to +5 samples. Looking at the maximum ccfmax and the corresponding pitch length of the 11 correlation values, we can adapt the pitch length kp on one hand and can stop the determination of further pitch periods when the maximum correlation ccfmax is below a threshold thres1. We choose a value of 0,8 for thres1.

The same procedure is applied to detect successive pitch periods towards higher sample indices starting again at sample index n_{mid}. Instead of calculating the correlation between two pitch periods backward in time we move forward in time.

We observed the determination of too long voiced segments with too many pitch periods in cases where two voiced phonemes consecutively occur. This is due to calculating only the similarity between each pair of successive pitch periods. But we do not take into account the similarity to the speech segment where we detected the maximum frame energy and where we started the processing. To avoid this misdetection of too many periods we calculated the similarity of each detected pitch period with the pitch period starting at sample index n_{mid} by calculating the correlation according to equation 1.

$$
ccf(p) = \sum_{k=0}^{kpitch_{est}-1} s(n_{mid}+k)\cdot s(pind(p)+k) \qquad (1)
$$

$$
with\ \ p = 1,\ldots, numberofpitchperiods
$$

Taking only these contiguous pitch periods with a correlation value ccf higher than a threshold thres2 we are able to avoid the detection of too long voiced segments. We choose a value of 0,6 for thres2.

Applying the mentioned processing we observed the identification of a fairly low number of pitch periods for some of the voiced phonemes due to the choice of the two thresholds thres1 and thres2. Thus, we iteratively repeated the whole processing by lowering the thresholds in steps of 0,1 until we detect a minimum number of pitch periods. We choose a minimum number of 8 for our first experiments and as minimum correlation for thres1 the value 0,5 and for thres2 0,3.

In the following, we estimate the spectral components of each pitch period. We take the samples of a single period and apply a DFT. The prerequisite of applying the DFT is the analysis of exactly one or more periods of a periodic signal. Otherwise, the DFT spectrum contains errors due to the leakage effect. We fulfill this prerequisite so that we do not need the weighting with a window function to reduce the errors due to the leakage effect. Each pitch spectrum is normalized to its energy because we are only interested in the spectral shape. We take the logarithm of the magnitude spectrum to take into account the nonlinear loudness perception of humans. Finally, a DCT (Discrete Cosine Transform) is applied to the logarithmic spectrum to get acoustic features that are independent of each other. Due to the mentioned energy normalization we ignore the cepstral coefficient C0. We choose the number of cepstral coefficients to the integer value of half of the number of spectral components. Due to the usually varying pitch length we get a sequence of spectra that contain a varying number of components. To generate a vector sequence with a fixed number of acoustic features per vector we apply a pitch

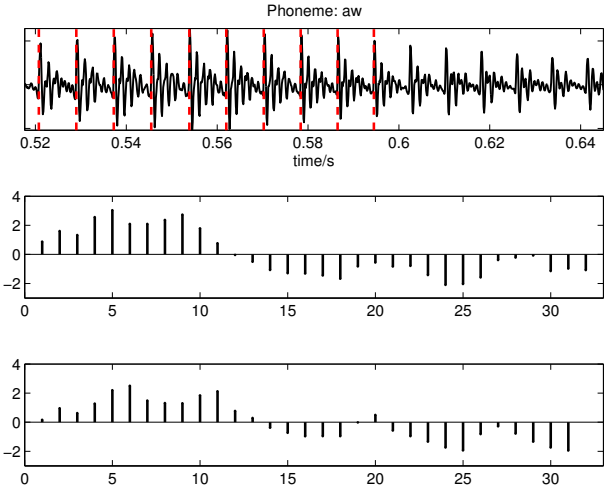

Figure 2: Detected pitch periods and pitch spectra.

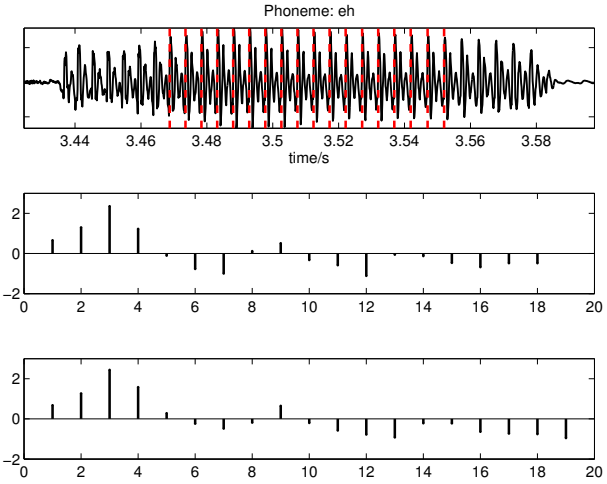

Figure 3: Detected pitch periods and pitch spectra.

dependent DCT that maps the varying number of spectral components to a fixed number of cepstral components. This procedure is clarified by looking at an example. Assuming the detection of pitch periods with a varying length between 78 and 86 samples, we get spectra with 78/2+1=40 up to 44 components. We take the minimum length of 40 to define the desired number of cepstral coefficients as half of this length. Then, we apply an individual DCT for each spectral length with the goal to calculate a fixed number of 20 cepstral coefficients.

Figures 2 and 3 contain two examples for the pitch synchronous analysis visualizing the detected periods for the phoneme "aw" of a male speaker and the phoneme "eh" of a female speaker. Furthermore, the logarithmic magnitude spectra of the first and the last detected period are plotted in the middle and in the lower part of the figures. Comparing the number of DFT components of the first and the last pitch period they differ in one component in both examples. All spectra show the typical formant characteristics of voiced phonemes.

3. Pitch Dependent Recognition

Applying the knowledge about a probably pitch dependent processing and recognition in the human auditory system we perform a pitch dependent recognition as it is shown in figure 4.

The pitch synchronous spectral analysis as described in the preceding section is split into two blocks. The pitch dependent DCT is separated because it can be used to create a desired number of cepstral coefficients. By splitting the whole pitch frequency range into 15 regions we define the number of cepstral coefficients dependent on the pitch region where the pitch of an individual phoneme segment falls into. We estimate the pitch length of all periods as part of the pitch synchronous analysis. As mentioned before we focus on the minimum length of all periods. This minimum pitch length is taken to assign the phoneme segment to one of the 15 pitch regions. Different approaches are possible to split the pitch range from about 70 up to 330 Hz into a certain number of regions. The linear splitting into the desired number of regions would be a possibility.

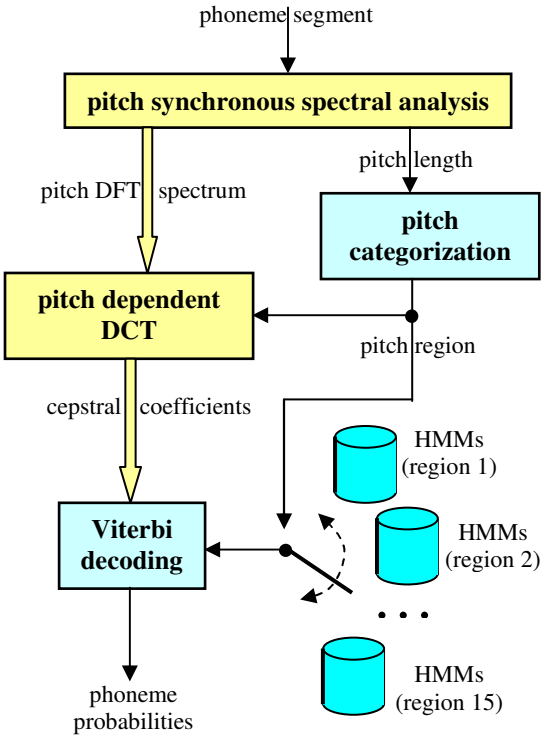

Figure 4: *Pitch dependent recognition.*

Another possibility would be the statistically based splitting to achieve the occurrence with equal probability in each region. Instead of the mentioned possibilities we looked at the number of cepstral coefficients that we create as final acoustic parameters. We split the pitch range into 15 regions by selecting a number of cepstral coefficients between 11 and 39 with a step size of 2. Thus, we split the pitch range as shown in table 2. We define fairly small regions for low pitch frequency and broader regions for high pitch frequency.

Number of cepstral coefficients	Pitch (F0) range
11	F0 > 320 Hz
13	276 Hz < F0 ≤ 320 Hz
15	242 Hz < F0 ≤ 276 Hz
…	…
37	99 Hz < F0 ≤ 104 Hz
39	F0 ≤ 99 Hz

Table 2. *Pitch frequency regions*

The assignment to one of the 15 pitch regions is used to define the pitch dependent DCT and especially to fix the number of cepstral coefficients for all periods of each phoneme segment. Furthermore, this pitch categorization is taken to select a corresponding set of HMMs. Each HMM set consists of 15 models for the 15 voiced phonemes. All HMMs of an individual set contain the feature characteristics within the corresponding pitch region. We choose a modeling with 3 states as shown in figure 5.

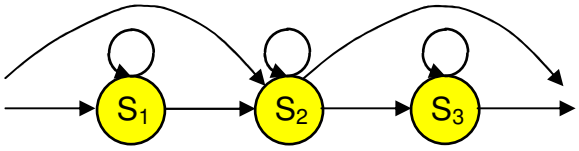

Figure 5: *HMM structure.*

Due to the detection of a varying number of pitch periods we define a transition to the middle state as possible entry into the model and a transition from the middle state as possible exit out of the model. Thus, we want to take into account the case where we detect only a low number of pitch periods that belong to the kernel part of a phoneme.
To estimate the HMM parameters we analyze the about 55000 voiced segments of the 4620 Timit training utterances according to the scheme shown in figure 4. Taking all vector sequences of a single phoneme that are assigned to a specific pitch region we apply the tools of HTK [14] to define the multi-variate Gaussian distributions of the 3 HMM states as well as the transition probabilities. We choose a modeling with a mixture of 4 Gaussian distributions.

4. Recognition Experiments

To verify the applicability und the usefulness of the pitch synchronous analysis and the pitch dependent recognition we ran a first set of recognition experiments. We looked at the about 55000 voiced segments of the 4620 Timit training utterances. The recognition task is the classification of each segment to one of the 15 phoneme classes.

With our first experiments we focus on the kernel part of each phoneme to reduce the influence of coarticulation effects due to different phonemes in the direct neighborhood. Therefore, we slightly modify the pitch synchronous analysis scheme as described before. We restrict the number of successive pitch periods to a maximum length of 45 ms. This is realized by selecting these pitch periods with highest correlation so that we do not exceed the mentioned length. We choose the period of 45 ms because this length corresponds to the length of 3 frames with our robust cepstral analysis scheme [4]. This analysis scheme consists of calculating 12 MFCCs and the frame energy as well as the corresponding Delta and Delta-Delta coefficients for segments of 25 ms and a window shift of 10 ms. Therefore, 3 frames correspond to 45 ms. We select the 3 frames by detecting this frame with highest energy. This is the same procedure that we take as starting point for the pitch synchronous analysis. Thus, we analyze with both techniques almost the same segment of each phoneme.

Restricting the pitch analysis to a maximum length of about 45 ms corresponds to the detection of about 4 periods for a pitch frequency of 100 Hz and of about 9 periods for a pitch frequency of 200 Hz. This does not mean that we are always able to detect this number of periods. We observe cases where we detect only two or even only a single period.

To perform the recognition with the 3 frames of the robust cepstral analysis we take the monophone HMMs whose parameters have also been trained from the corresponding features of the 55000 segments. Each HMM consists of 3 states where the occurrence of the 39 acoustic features is modeled as a mixture of 16 Gaussian distributions in each state. A usual training procedure [13] with HTK is applied to estimate the HMM parameters of all 39 phonemes. For our experiment we take only the middle states of the 15 voiced phonemes and use them as GMMs (Gaussian Mixture Models). We calculate the mean of each acoustic parameter over the 3 frames. These means are taken to calculate the probabilities to generate them with the 15 GMMs. Phoneme recognition rates are listed in table 3 for the 55000 segments. Besides the results on the clean data we present also results for four noisy versions of the 4620 Timit utterances. We created these noisy versions with the FaNT tool [15] by randomly adding different segments of several noise recordings at different SNRs. Two noise scenarios are investigated. These are the communication inside a car (car5db and car0db) and in a noisy room like a restaurant or an exhibition hall (int5db and int0db) at SNRs of 5 and 0 dB.

acoustic condition				
clean	car5dB	car0dB	int5db	int0db
43,4 %	39,9 %	35,4 %	35,1 %	27,8 %

Table 3. *Phoneme recognition rates (cepstral features)*

We observe fairly low recognition rates due to the fact that we focus on the kernel part of each phoneme by selecting only 3 frames. The degradation due to the noise is not that high where the robust analysis scheme might be the reason for this.

Applying alternatively the pitch synchronous analysis and the pitch dependent recognition we achieve the phoneme recognition rates listed in table in 4.

acoustic condition				
clean	car5dB	Car0dB	int5db	int0db
59,1 %	42,2 %	35,3 %	39,1 %	31,0 %

Table 4. *Phoneme recognition rates (pitch synchronous features)*

The recognition rate for the clean data is considerably higher than with the usual cepstral features. The degradation for the noisy data is higher in comparison to the robust cepstral features. This is due to the fact that we did not design the pitch synchronous analysis to be robust against additive noise so far.

5. Conclusions

We present the idea of a pitch synchronous analysis and a pitch dependent recognition derived from the observation and the knowledge that humans use the pitch information as important feature for their communication. We could show with a first simple recognition experiment that there seems to be a potential to improve the recognition accuracy with this approach. The presented details of the processing may only be considered as an intermediate result. For example, we intend to investigate alternative approaches for detecting pitch periods and apply techniques like PCA (principal component analysis) and LDA (linear discriminant analysis) to derive characteristic features from the pitch spectra. Our focus will be on increasing the robustness with respect to noise and other distortion effects.

6. Acknowledgements

The author would like to thank the German funding organization DFG (Deutsche Forschungsgemeinschaft / German research community) for supporting this work.

7. References

[1] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm”, ETSI document ES 202 050 v1.1.3 (2003-11), Nov. 2003.

[2] M.J.F. Gales, S. Young, “Robust speech recognition in additive and convolutional noise using parallel model combination”, Computer, Speech and Language, Vol. 9, 1995.

[3] C.J. Leggeter, P.C. Woodland, “Maximum Likelihood Linear Regression for speaker adaptation of continuous density Hidden Markov Models”, Computer Speech and Language, Vol.9, 1995

[4] H.G. Hirsch, A. Kitzig, “Robust speech recognition by combining a robust feature extraction with an adaptation of HMMs”, ITG symposium Speech Communication, 2010.

[5] H.G. Hirsch, F. Finster, “A new approach for the adaptation of HMMs to reverberation and background noise”, Speech Communication, Vol.50, pp. 244-263, 2008.

[6] V.W. Zue, “The use of speech knowledge in automatic speech recognition”, Proc. of IEEE, Vol.73, 1985.

[7] M.P. Cooke, P.D. Green: “Robust automatic speech recognition with missing and unreliable acoustic data”, Speech Communication, Vol.34, 2001.

[8] P. Zhan, A. Waibel, “Vocal Tract Length Normalization for large vocabulary continous speech recognition”, Technical report CMU-CS-97-148, Carnegie Mellon University, 1997.

[9] Y. Medan, E. Yair, “Pitch synchronous analysis scheme for voiced speech”, IEEE ASSP, Vol. 37, No. 9, 1989.

[10] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, C.A. McGonegal, “A comparative performance study of several pitch detection algorithms”, IEEE ASSP, Vol.24, No. 5, 1976.

[11] W.M. Fisher, G.R. Doddington, K.M. Goudie-Marshall, “The DARPA speech recognition research database: specification and status”, Proc. Darpa workshop, pp. 93-99, 1986.

[12] “The CMU pronouncing dictionary”, available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[13] K. Vertanen: “Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments”, Technical report, University of Cambridge, 2006.

[14] S. Young et al., “The HTK book (for version 3.3)”, available at http://htk.eng.cam.ac.uk , 2005.

[15] “Filtering and Noise adding Tool - FaNT”, available at http://dnt.kr.hsnr.de/ in the download section