ROBUST VOICED-UNVOICED DECISION ASSOCIATED TO CONTINUOUS PITCH TRACKING IN NOISY TELEPHONE SPEECH

Mijail Arcienega and Andrzej Drygajlo

Swiss Federal Institute of Technology, Lausanne
Signal Processing Institute
{mijail.arcienega, andrzej.drygajlo}@epfl.ch

ABSTRACT

In this paper, a new algorithm for a robust voiced-unvoiced classification of frames in telephone speech is presented. This algorithm is associated to a continuous pitch contour extraction procedure based on a discrete logarithmic Fourier transform and a dynamic programming search as introduced in [1]. While the pitch extraction procedure is forced to find a pitch value for every frame, even in unvoiced regions, the algorithm presented in this paper is capable of making a robust voiced-unvoiced decision by using multiple features and two-state ergodic hidden Markov modeling (HMM).

1. INTRODUCTION

The pitch is one of the most frequently used prosodic features. Prosodic features are known to be less affected by noise than the spectral envelope features. The pitch was already used in the pioneer speaker identification experiments by Atal [2]. Today, it is widely used in speech synthesis and coding systems, as well as a useful prosodic feature for speech and speaker recognition systems [3], [4], [5].

Recently, a pitch-dependent Gaussian mixture model (GMM) for speaker recognition systems has been developed in our laboratory [3]. This model aims at exploiting the difference between voiced and unvoiced parts of speech by modeling these two parts separately. However, the performance of the system strongly relies on a proper discrimination of the voiced and unvoiced frames.

Although several algorithms have been developed for extracting the pitch from the speech waveform in a reliable way [6], there is still a need for their improvement in very noisy conditions. The pitch does not exist in unvoiced speech regions and a pitch extraction algorithm must be also capable of making a correct voiced-unvoiced decision in presence of noise. In this paper, an improved continuous pitch extraction algorithm, based on previous works by Wang and Senef [1], originally developed for telephone quality speech, is modified and completed with a multi-feature classification of the voicing status. The algorithm for extracting the continuous pitch contour uses the discrete logarithmic Fourier transform (DLFT) and a dynamic programming (DP) search for tracking the best pitch path. The classification into voiced and unvoiced speech frames is made by using a three-feature vector ψ composed by an entropy measure of the spectrum, the energy and a correlation measure defined by the continuous pitch extraction algorithm.

2. CONTINUOUS PITCH CONTOUR TRACKING ALGORITHM

We present here a description of the improved algorithm for the continuous pitch contour extraction. Major changes with reference to the original algorithm are introduced in section 2.2.

2.1. Discrete Logarithmic Fourier Transform

Given a window of speech signal \( x(n), n = 0, 1, ..., N_w - 1 \), the DLFT for the frequency band \([f_1, f_2]\) is computed as:

\[
X_i = \frac{1}{N_w} \sum_{n=0}^{N_w-1} x(n) \cdot e^{-j \omega_i n} \quad (i = 0, 1, ..., N - 1),
\]

where

\[
w_i = 2 \pi \cdot e^{(\ln(f_1) + i \cdot dlnf) \cdot T_s} \quad (i = 0, 1, ..., N - 1),
\]

\[dlnf = \frac{\ln(f_2) - \ln(f_1)}{N - 1},\]

\(N\) is the size of the DLFT, \(T_s\) is the sampling period of the waveform.

In a logarithmic scale, the fundamental frequency \(F_0\) and its harmonics appear as peaks at \(\ln(F_0), \ln(F_0) + \ln 2, \ln(F_0) + \ln 3, ..., \) etc. The position of the first peak depends on \(F_0\), but subsequent peaks are spaced in the same way (\(\ln 2, \ln 3, \ldots\)) relative to this first peak. In order to find the \(F_0\) value, the cross correlation between the DLFT of the speech window and the DLFT of a k-pulse template is calculated. Thanks to this method, even if \(F_0\) is missing, its value can be still retrieved. The position of the maximum of this cross correlation function gives the difference between the fundamentals of the template and of the speech frame, \(k\) being the number of harmonics taken into account.

2.2. Correlation Functions

A Hamming windowed 200 Hz impulse train is used as a template. The speech signal is also windowed with a Hamming window. After the DLFT is applied to both signals, they are \(\mu\)-law compressed in order to rehearse the peaks in the spectrum. Let \(T(n)\) be the normalized template to have unit energy and \(X_i(n)\) the \(\mu\)-law converted DLFT for the \(i^{th}\) frame. The normalized correlation between \(T(n)\) and \(X_i(n)\) can be written as:

\[
R_{TX_i}(n) = \frac{\sum_i T(i) \cdot X_i(i-n)}{\sqrt{\sum_i T^2(i) \cdot \sum_i X_i^2(i)}} \quad (N_L < n < N_H)
\]
where the range for the calculations $[N_L, N_H]$ is determined by the $F_0$ range to be analyzed.

In order to provide a constraint for the DP search, the cross-frame correlation is also calculated:

$$R_{X_0X_{i-1}}(n) = \frac{\sum X_i(i) \cdot X_{i-1}(i-n)}{\sqrt{\sum X_i(i)^2 \sum X_{i-1}(i)^2}} \quad (|n| < N) \quad (3)$$

The maximum of $R_{X_0X_{i-1}}(n)$ provides a measure of the variation of $F_i$, $\delta \ln(F_i)$. Bounds in the DP are weighted by this function.

It has been seen that the DP search (described in next section) originally applied to the $R_{TX_i}(n)$ provides correct results in most cases, nevertheless, in some cases, and particularly when the level of the noise increases, the peaks (that we want to track) at time $i$ do not correspond anymore to the maximum in this function. This maximum (now at the top of a smooth lobe) corresponds to the position of the harmonic with highest energy. This phenomenon is increased by the use of a logarithmic frequency scale and the DP search therefore provides an incorrect pitch value. In order to correct this problem, a high pass filter $h(n)$ is applied to the $R_{TX_i}(n)$ function, this high pass filter catches the target peak (which becomes the maximum) and eliminates the “false” maximum at the top of the smooth lobe.

$$R^f_{TX_i}(n) = h(n) \ast R_{TX_i}(n) \quad (4)$$

where $h(n)$ is a 32th order linear high-pass filter with a cut-off frequency equal to $\frac{1}{10} f_N$.

This additional step has no influence on the amount of total calculations, indeed, the template function can be modified in order to take into account the effects of this filter.

2.3. Dynamic Programing Search

The tracking of the pitch contour can be formulated as a problem of dynamic programming where the goal is to find the best path under certain constraints. This path is to be found by choosing the best $n$ for each $R^f_{TX_i}(n)$, $t = 1, 2, ..., T$, where $T$ is the number of frames in the speech sample. The target function that determines the best path is defined as:

$$S_t(i) = \begin{cases} \max_j \left\{ S_{t-1}(j) \cdot R_{X_0X_{i-1}}(i-j) \right\} & (t > 0) \\ R^f_{TX_i}(i) & (t = 1) \end{cases} \quad (5)$$

This target function ensures a very smooth pitch contour. The score at time $t$ for node $i$ is given by the template-frame correlation at $i$ plus the best past score weighted by the cross-frame correlation. Once the best score at time $T$ is found, backtracking is performed in order to find the best path.

Figure 1 shows the path that the algorithm is capable of tracking for various levels of noise. The tracked path is plotted over the spectrum of the utterance in order to show the effectiveness of the algorithm. The spectrum presents peaks (harmonics) in voiced frames while in unvoiced frames these peaks do not exist. We can see that the pitch values for voiced frames are correctly estimated for SNRs up to SNR=5[dB]. However, in unvoiced frames, where correlation is low, the estimated value of pitch has greater variability.

3. ENERGY AND ENTROPY MEASURE OF THE SPECTRUM

The energy of a frame is a reliable indicator of its voicing status in clean conditions. When the SNR is very low, the noise completely masks unvoiced frames and their energy is comparable to the energy of voiced frames. Spectrograms of speech in noisy environments show that voiced frames present a structure that is not present in unvoiced frames. The Shannon’s entropy is a suitable measure for the level of organization of such a structure present in voiced frames.

The measure of entropy in the spectral energy domain [7] is defined as:

$$H(|Y(f, t)|^2) = -\sum_{j=1}^F P(|Y(f, t)|^2) \cdot \log(P(|Y(f, t)|^2)) \quad (6)$$

where

$$P(|Y(f, t)|^2) = \frac{|Y(f, t)|^2}{\sum_{t=1}^T |Y(f, t)|^2}$$

is the estimated probability of the frequency $f$ in the magnitude spectrum at frame $t$.

4. THE VOICED-UNVOICED DECISION

Since the DP is forced to find a pitch value for every frame, even in the unvoiced regions, a parallel procedure is necessary for determining whether a frame is voiced or not. Several methods can be
employed for this task and most of them consist of applying a thresholding to the maximum of the correlation function: frames that present a maximum above the threshold are considered as voiced and the remaining are considered as unvoiced. This way of proceeding has all the disadvantages of fixed a priori threshold. Since the environment never remains the same, this threshold needs to be re-estimated for each specific case. An attempt to improve the decision was made in [1] where both the maximum of the correlation function $R_{TX}^{\text{DP}}(n_{\text{MAX}})$ and the minimum of $R_{X_tX_{t-1}}(n_{\text{MIN}})$ were used. Nevertheless, the $R_{X_tX_{t-1}}(n)$ function is strongly influenced by noise and does not provide a good discrimination at low SNRs.

In order to make a robust decision of voiced and unvoiced zones, several features are represented together in a vector $\psi$. This vector includes the entropy measure of the spectrum, the value of $R_{TX}^{\text{DP}}(n_{\text{DP}})$ for the tracked $n_{\text{DP}}$ at time $t$ and the energy. These features are plotted in Figure 2.

It is important to outline that $R_{TX}^{\text{DP}}(n_{\text{DP}})$ is not always equal to $R_{TX}^{\text{DP}}(n_{\text{MAX}})$, indeed the value of $R_{TX}^{\text{DP}}(n_{\text{DP}})$ being associated to the constrained tracked path is more discriminative.

While $R_{TX}^{\text{DP}}(n_{\text{DP}})$ and the energy $E$ present high values for voiced zones, the corresponding entropy is low. Voiced and unvoiced zones can be determined in a robust way by modeling this bi-modal behavior of $\psi$ with a two-Gaussian mixtures model.

4.1. Two-Gaussian Mixtures Model for the Voiced-Unvoiced Decision

A two-Gaussian mixtures model can be represented by

$$\lambda = \{ w_1, w_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2 \}$$

where $w_i$ are the weights of each mixture, $\mu_i$ are two 3-component vectors representing the means and $\Sigma_i$ representing the diagonal covariance matrices. We decided to use a diagonal covariance matrix after observing the diagonal predominance of the covariance matrix of the set of feature vectors.

Once the expectation-maximization (EM) algorithm has been performed in order to fit the distribution of $\psi$, the voiced-unvoiced status needs to be associated to each one of the mixtures. In order to do this, one can compare the second component (that belongs to the entropy) in the mean vectors $\mu_i$. If $\mu_2(2) < \mu_4(2)$ then the first mixture has been associated to the voiced frames and the second to the unvoiced frames; otherwise the opposite is true.

For determining whether or not a frame is voiced, the weighted likelihoods associated to each mixture are compared and the frame is then associated to the mixture (and therefore to a voicing state) with highest likelihood. This is the mechanism employed by a Bayesian classifier.

4.2. Two-State HMM for the Voiced-Unvoiced Decision

Although classifying the frames as voiced or unvoiced using two Gaussian mixtures performs well, some minor misclassification errors may occur in certain circumstances. In particular, an isolated frame can be classified as voiced in the middle of an unvoiced zone. This case is not realistic in speech when windows of about 50ms are used for the analysis. Voiced (and unvoiced) zones have durations greater than one frame. By using a two-state ergodic HMM, which models the transitions between voiced and unvoiced states, as well as the probabilities of remaining in the same (voiced-unvoiced) state, we can avoid this problem.

The two-state ergodic HMM for modeling the states can be trained by applying the Baum-Welch algorithm using the feature vectors $\psi$. As for any HMM, the main problem is to fix the initial parameters of the model.

These initial parameters can be calculated from the classification made by the two-Gaussian mixtures model explained before. In this way, the initialization is very close to the final solution and the Baum-Welch algorithm converges in a few iterations. The model of $\psi$ is then

$$\lambda^\psi = \{ A, \mu_1, \Sigma_1 \}$$

where $i = 1, 2$ and $A$ is the transition probabilities matrix.

In order to determine the sequence of states $S_{\psi t}$ for a sequence of vectors $O_{\psi t}$, two methods can be employed:

- to use $\gamma^\psi_t(i)$ for individually attributing the most likely state $s_{\psi t}$ to the frame at time $t$ as
  $$s_{\psi t} = \arg \max_{i=1,2} \gamma^\psi_t(i) \quad t = 1, \ldots, T,$$
  $$\gamma^\psi_t(i) \text{ being defined in the Baum-Welch algorithm as}$$
  $$\gamma^\psi_t(i) = p\left( s_{\psi t} = i | O_{\psi t}, \lambda^\psi \right)$$

- to find the best state sequence, i.e. to maximize $p(S_{\psi t} | O_{\psi t}, \lambda^\psi)$. This can be done by applying the Viterbi algorithm. The advantage of using this method is to avoid glitches that may occur in the previous method. This method was used in our speaker verifications experiments using the pitch-dependent GMMs [5].
The best sequence of states $S_\varrho$ for the pitch is equal to the best sequence of states $S_\psi$ for $\psi$.

Telephone quality speech of the Switchboard database was used in our experiments. The quality of the speech signals of this database is considered as the reference. White noise is then added to the telephone quality speech. SNR is calculated as the overall ratio of the reference signal energy and the noise energy in an utterance.

Figure 3 shows the decision made by using this technique on the pitch values of Figure 1.

**Figure 3.** Voiced-unvoiced decision using a two-state HMM for $\psi$

5. CONCLUSIONS

Tests were performed on telephone quality speech where the fundamental frequency $F_0$ is often missing. The improved DP search is capable of tracking a continuous pitch contour and retrieving this missing fundamental frequency even at low SNRs. The voiced-unvoiced decision made on speech of the Switchboard database was considered as the reference (telephone quality). White noise was added then to verify the robustness of our voiced-unvoiced decision method. As shown in Figure 3, this discrimination presents only a few differences when comparing with the reference up to a SNR=5[dB].

It is important to outline that the algorithm presented in this paper does not need any a priori fixed threshold for the classification of the frames. Indeed, the algorithm is independent on the level of noise. Moreover, by using three features instead of one, the voiced-unvoiced decision becomes less affected by noise.

Although the training and testing stages, when using a two-Gaussian mixtures, demand less computational load, the two-state HMM version provides improved classification by using the transition probabilities which model the average voiced and unvoiced state durations.

6. REFERENCES


