



Adaptive weighting of microphone arrays for distant-talking F0 and voiced/unvoiced estimation

Federico Flego¹, Christian Zieger², Maurizio Omologo²

¹ Radio Trevisan, Via Pietraferrata 9, 34147 Trieste, Italy

² FBK-irst, via Sommarive 18, 38050 Povo, Trento, Italy

federico.flego@radiotrevisan.com, zieger@itc.it, omologo@itc.it

Abstract

This paper introduces a new technique of multi-microphone processing which aims to provide features for the extraction of fundamental frequency and for the classification of voiced/unvoiced segments in distant-talking speech. A multi-channel periodicity function (MPF) is derived from an adaptive weighting of normalized and compressed magnitude spectra. This function highlights periodic clues of the given speech signals, even under noisy and reverberant conditions. The resulting MPF features are then exploited for voiced/unvoiced classification based on Hidden Markov Models. Experiments, conducted both on simulated data and on real seminar recordings based on a network of reversed T-shaped arrays, showed the robustness of the proposed technique.

Index Terms: microphone arrays, pitch extraction, voiced/unvoiced speech classification, distributed microphone network, acoustic features.

1. Introduction

The use of a Distributed Microphone Network (DMN), that consists in a generic set of microphone arrays localized in space without any specific geometry, is emerging in acoustic scene analysis. Some of the related applications, as investigated in the CHIL project (see <http://chil.server.de>), are distant-talking speech recognition, speaker localization, speaker identification, acoustic event classification, speech activity detection. For most of those tasks, deriving reliable acoustic features from the given far-microphones represents an ambitious objective.

This work focuses on the problem of estimating robustly the fundamental frequency F0 and of classifying voiced/unvoiced speech from the variety of signals recorded through a DMN. In a reverberant and noisy environment speech signals acquired by far microphones are severely degraded. The distortion depends on spatial relationships among microphones and speaker (distance and orientation), as well as on the acoustic scene (obstacles between speaker and microphones, diffused and/or localized noise sources).

There are several possible methods to take advantage of the DMN to estimate F0 or classify voiced speech. For example a possible approach is to derive an estimation independently for each microphone signal and applying then majority vote or other fusion based methods. Another way is to extend to the multi-microphone case a paradigm that works for a single microphone close-talking case.

In previous works [1, 2], an algorithm based on a Multi-microphone Periodicity Function (MPF) was introduced and

This work was partially funded by the European Community, under the CHIL and DICIT projects.

compared to the multi-microphone Weighted Autocorrelation (WAUTO) [3] and to a multi-microphone extension of the YIN algorithm [4]. In particular, the resulting multi-microphone technique offered the advantage of obtaining better performance than single microphone based processing, without any assumption or knowledge about the position of the microphones as well as of the talker. Experiments conducted in a noisy and reverberant environment showed the effectiveness of that technique.

This paper introduces a change in the computation of the channel reliability degree to the framework already described in [2], which allows a more effective discriminative dynamic selection of the most useful microphones.

The work also deals with the problem of voiced/unvoiced speech classification. There are several approaches described in the literature for this problem and the most reliable ones adopt a statistical approach exploiting a combination of features to well discriminate the two classes [5, 6, 7]. However in distant-talking under highly reverberant conditions, V/UV classification becomes a surpassingly difficult task. To tackle it, we propose a Hidden Markov Model (HMM) based classifier which uses a three dimensional feature vector as acoustic observation. Two features derive from the MPF and one is energy based. Experimental results reported in the following show the advantages of the here proposed technique.

The paper is organized as follows: Section 2 introduces MPF based F0 extraction algorithm; Section 3 describes the proposed voiced/unvoiced classification system; Sections 4 and 5 report on the experimental set-up and on evaluation criteria; Section 6 describes the experimental results; the last section draws some conclusions.

2. MPF based F0 extraction

Given the above mentioned DMN context, the signals acquired at each microphone are differently distorted due to the reverberation effect which may affect the spectral peaks relative to F0 and its harmonics, modifying their magnitude and shifting their frequency locations (see Figure 1). A way to overcome these distortion effects is to exploit the redundancy offered by the DMN and to derive a common harmonic structure from the different spectra by weighting more the most reliable channels.

Given a DMN consisting of M microphones, a downsampled version of the speech signal recorded at the i -th microphone is windowed with a window function of length L_w to produce for each analysis frame the vector $\mathbf{s}_i = [s_i(1) \dots s_i(L_f)]$. An L_f -points FFT is then computed and processed as follows:

$$S_i(k) = |\text{FFT}\{\mathbf{s}_i\}(k)|^\gamma \quad (1)$$

being k the frequency bin and γ a spectral compression fac-

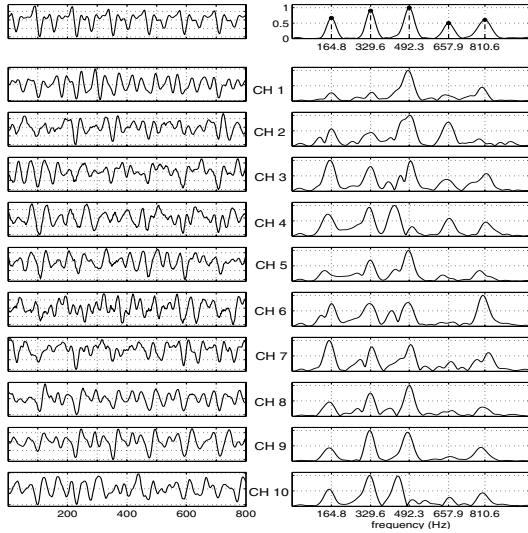


Figure 1: Example of speech signals and their spectra for a close-talk and for ten distant-talking acquisitions provided in a reverberant room with $T_{60} = 0.45 \text{ sec}$, defined as the time taken for the sound pressure level to decrease by 60 dB due to absorption in the walls.

tor ($\gamma = 2$ provides the power spectrum). Zero-padding is applied to \mathbf{s}_i if $L_f > L_w$. Now, we consider the normalized compressed magnitude spectrum $X_i(k)$ for $k = 1, \dots, K$ ($K = L_f/2 + 1$) defined as

$$X_i(k) = \frac{S_i(k)}{\sqrt{\sum_{l=1}^K S_i^2(l)}} \quad (2)$$

Note that according to experimental evidence the proposed normalization outperforms the maximum based one investigated in [1]. Then a weighted sum of normalized functions $X_i(k)$ is computed:

$$\bar{S}(k) = \sum_{i=1}^M c_i X_i(k), \quad 1 \leq k \leq K. \quad (3)$$

where c_i are real weights. Next, IFFT is applied to obtain the Multi-microphone Periodicity Function $MPF(l)$ in the lag-domain

$$MPF(l) = \text{IFFT}\{\bar{S}(1), \dots, \bar{S}(K), \bar{S}(K-1), \dots, \bar{S}(2)\}(l) \quad (4)$$

in a similar manner as described in [8]. The lag value at which a maximum of $MPF(l)$ is found can be considered as an approximation of the fundamental frequency period T_0 estimate. To improve lag resolution, an interpolation of a factor Q is applied to $MPF(l)$ to obtain $MPF(l')$ and derive the T_0 estimate as:

$$T_0 = \arg \max_{l'} \{MPF(l')\}, \quad T_{\min} \leq l' \leq T_{\max}, \quad (5)$$

where T_{\min} and T_{\max} are the minimum and maximum fundamental frequency periods, respectively.

To assign the weight values, c_i , first a reference smoothed spectrum $P(k)$ is estimated as a product of channel magnitude spectra:

$$P(k) = \prod_{i=1}^M X_i(k). \quad (6)$$

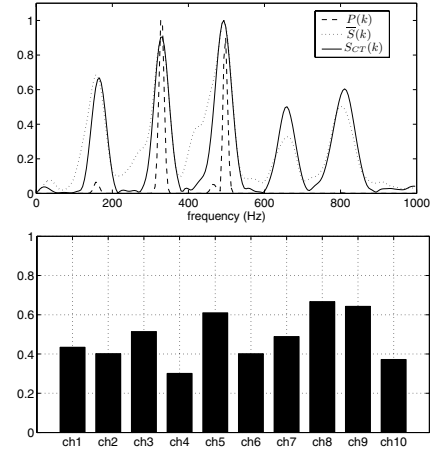


Figure 2: In the top figure, $P(k)$, $\bar{S}(k)$ and the close talk spectrum $S_{CT}(k)$ are compared for the example reported in Figure 1. In the bottom figure, the c_i values used in the computation of $\bar{S}(k)$ are plotted.

In this way, $P(k)$ will retain information common to the different channels while rejecting interference represented by frequency patterns not common to all channels (see Figure 2). Then each weight c_i is derived based on the inner product between vectors applied to $\mathbf{X}_i = [X_i(1) \dots X_i(K)]$ and $\mathbf{P} = [P(1) \dots P(K)]$ after having subtracted the mean value

$$c_i = \frac{(\mathbf{P} - m_p) \cdot (\mathbf{X}_i - m_i)}{\|\mathbf{P} - m_p\| \|\mathbf{X}_i - m_i\|} \quad (7)$$

where $m_p = 1/K \sum_{k=1}^K P(k)$ and $m_i = 1/K \sum_{k=1}^K X_i(k)$. Coefficients c_i represent the cosine of the angle between $(\mathbf{P} - m_p)$ and $(\mathbf{X}_i - m_i)$ and will thus range from -1 to 1. They have been introduced to represent the reliability degree of each component $X_i(k)$, which may depend on the speaker position, head orientation or on the presence of other noisy sources. Reliable channels are characterized by high values of c_i ; channels with $c_i < 0$ (completely unreliable) are unlikely and are not considered in the computation of $\bar{S}(k)$ by setting c_i to zero.

In Figure 2 $P(k)$, $\bar{S}(k)$, the close talk spectrum $S_{CT}(k)$ and the reliability degree of each channel are shown for the example reported in Figure 1. Observing Figures 1 and 2 one can note the good agreement between the reliability channel and the quality of each spectrum. The most reliable channels are 5-8-9, while the least reliable ones are channels 4-6-10.

3. Voiced/unvoiced speech classification

The voiced/unvoiced speech classifier here proposed is based on HMMs. The three dimensional feature vector is defined as follows: $\Phi = [\Phi_1, \Phi_2, \Phi_3]$. The first two components derive from the MPF and are defined as:

$$\Phi_1 = MPF(T_0)/MPF(0), \quad \Phi_2 = MPF(1)/MPF(0). \quad (8)$$

Φ_1 is the ratio between the MPF peak value computed at T_0 and the value of MPF in 0; Φ_2 is the ratio between the value of MPF for lag 1 and the value of MPF in 0. Basing on the analogy of the MPF with the autocorrelation, high values of these features correspond to voiced frames. The third feature is energy based and is given by the logarithmic signal energy averaged among

all microphones:

$$\Phi_3 = 1/M \sum_{i=1}^M \log \left(\frac{\|s_i\|^2}{L_w} \right). \quad (9)$$

Frames with high values of Φ_3 correspond to voiced segments. Usefulness of Φ_3 is more evident under less adverse conditions.

This combination of features has been selected among others on the basis of a preliminary experimental investigation. Voiced and unvoiced speech are then modeled by two HMMs, each based on a two-state left-right model. The output probability density of the emitting states are described by a three dimensional Gaussian with diagonal covariance matrix. Mean and covariance matrix of the output probability densities and transition probabilities were estimated through the Baum-Welch algorithm. A common model was trained on all the speakers. A loop grammar with equiprobable transitions was adopted in the following experiments.

4. Experimental setup

Two databases are used to test the proposed techniques. The first one is a real spontaneous speech corpus collected under the CHIL project¹, which consists of 13 recordings, each about 5 minutes long, extracted from seminar sessions held at the Karlsruhe University.

Each speaker, wore a ‘‘Countryman E6’’ close-talking microphone, to capture a noise-free, non reverberant speech signal. The DMN layout, as shown in Figure 3, includes four reversed ‘‘T’’-shaped microphone arrays, each consisting of 4 microphones. Inter-microphone spacing is 20 cm and 30 cm along the horizontal and vertical directions, respectively. Speech sequences were recorded with 44.1 kHz sampling rate and 16 bit resolution.

The room is 7.10 m \times 5.90 m wide and the ceiling height is 3 m. Reverberation time was $T_{60} \simeq 0.45s$.

The second database is obtained by simulating a speech source and a noise source given in the room layout previously described. The speaker is positioned 1 meter far from array A. For the noise source two positions are considered: 1 meter far from array C (position 1) and in the center of the room (position 2). The speech source is characterized by a cardioid directivity pattern (simulated by a modified version of the image method described in [9]) and is oriented towards array D. The noise source is assumed as omnidirectional. The speaker is simulated using the close-talk signal belonging to the real corpus formerly described. First the noise signal is scaled on the basis of the imposed SNR; then speech and noise are filtered with the impulse response generated for each of the 16 microphones. The final signal acquired at each microphone is obtained by adding the filtered noise to the filtered close-talk signal.

5. Evaluation criteria

To evaluate the proposed F0 estimation algorithm, reliable and very precise ‘‘ground-truth’’ pitch estimates were derived as reported in [1]. In general this was accomplished applying three existing and well established pitch extractors, namely Praat, SFS and WaveSurfer, to the close-talk recordings and combining their pitch information to produce the final reference. A frequently used method to compare the performance between

¹A description of the used speech corpora can be found at <http://chil.server.de>, <http://www.nist.gov/speech> and <http://www.clear-evaluation.org>.

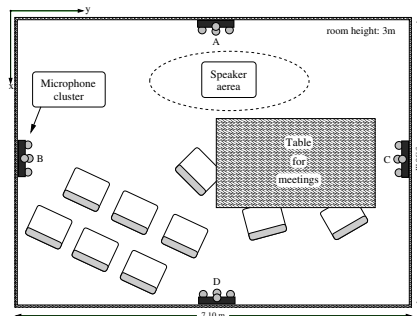


Figure 3: The CHIL room layout of Karlsruhe University. The dimension are 7.10 x 5.90 x 3 meters and the reverberation time is 0.45 seconds.

different F0 estimation algorithms is to compute the Gross Error Rate (GER). This is calculated considering the number of estimates which differ by more than a certain percentage from the reference values. In this work thresholds was set to 20% and 5% for the GER estimation, indicated with GER(20) and GER(5), respectively.

Voiced/unvoiced speech references were obtained from the set of voiced/unvoiced speech outputs produced by the same three reference pitch extractors. The final reference combines the information given by the three outputs by labeling each frame with voiced or unvoiced if at least two outputs are labeled as voiced or unvoiced respectively. For a given frame, an error occurred when, comparing the tested system output with the produced reference, there was a mismatch. The system performance are measured in terms of the total voicing error VUV_{error} , given by the ratio between the total number of errors and the total number of analyzed frames.

6. Results

The experiments were based on the following settings: the analysis frame is 60 ms and the analysis step is 10 ms; the downsampling factor is 9; the window type is Hanning; the compression factor γ is set to 0.6; the FFT point number is 1024; and the interpolation factor Q is 4. All of those choices are based on optimal assignments investigated in previous works [1, 2].

F0 estimation through MPF is applied to the real database with or without considering the channel reliability. The results obtained are GER(20) = 1.7% and GER(5) = 7% for the former case and GER(20) = 1.7% and GER(5) = 7.15% for the latter one. As the database is collected in diffuse noise conditions, with SNR of about 10 dB, there is only a very little improvement in terms of GER(5) when the channel reliability is considered. To show the effectiveness of coefficients c_i in presence of a coherent noise source, experiments were conducted on the simulated database. Different types of noises (white, pink, etc...), with omnidirectional or unidirectional radiation, were considered and all demonstrated the effectiveness of this technique. For sake of simplicity, in this work only the results in presence of an omnidirectional pink noise source are reported. In Figure 4 the GER(20) and GER(5) are compared for two noise positions, with or without the use of the weights. As expected, position 1 provides best results when compared with position 2. For position 1 arrays A-B-D are characterized by a higher SNR than array C, the nearest to the noise source. For this reason the estimated weights, c_i , for array C have values on average less than the ones of the other arrays. For position 2 all arrays are characterized by about the same SNR, i.e. there

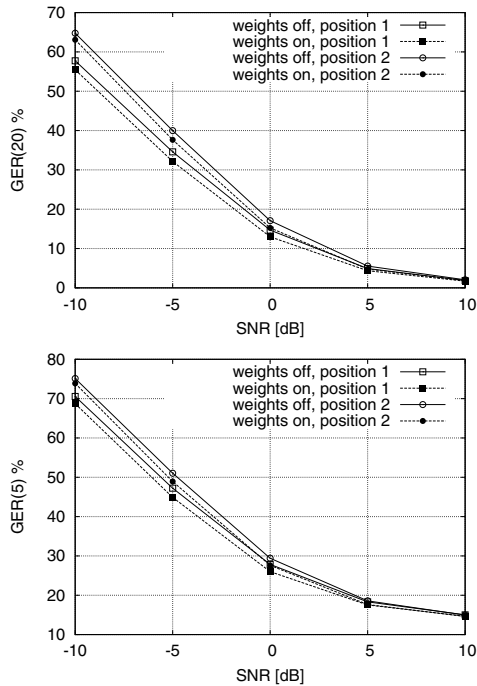


Figure 4: GER with threshold at 20% and 5% versus SNR.

are not more reliable microphones, which nevertheless allow a more accurate estimation of F0. Independently from the noise position the use of the channel reliability gives benefits to the system performance. Let us notice that here the performance improvement is evident only in conditions of very low SNR, in fact when the SNR increases all curves converge to the same value. However, in the general case one can have a higher reverberation time and a less favorable situation in terms of position of the source with respect to microphones. For all these cases a blind application of the proposed technique is always advantageous. For the voice classification experiments a lower down-sampling factor is used to obtain a working sampling frequency of 14700 Hz and no interpolation is applied. The other system parameters are the same used for F0 estimation. The choice of a higher sampling frequency was derived from preliminary experiments where system performance was compared varying the sampling frequency. The voice/unvoiced speech classifier was tested with the real speech corpora. The training data consisted of the first minute of each meeting for a total of 13 minutes. Figure 5 reports on the $VUV_{error}\%$ results obtained applying the voiced/unvoiced classifier on the close-talk signal, on each far microphone (MPF with only one microphone input) and on the DMN. The x-axis represents the microphone index (for the single microphone case) while the y-axis represents the $VUV_{error}\%$. Microphones 0-3 belong to array A, 4-7 to array B, 8-11 to array C and 12-15 to array D. Experiments show that the proposed voiced/unvoiced classifier yields better results than the single microphone case, although it still differs from the reference close-talking case of about 6%.

7. Conclusions

In this paper we proposed a new method for estimating the fundamental frequency and for classification of voiced/unvoiced speech in a reverberant and noisy environment exploiting the redundancy of the DMN. The MPF was exploited to combine the information given by each microphone considering the

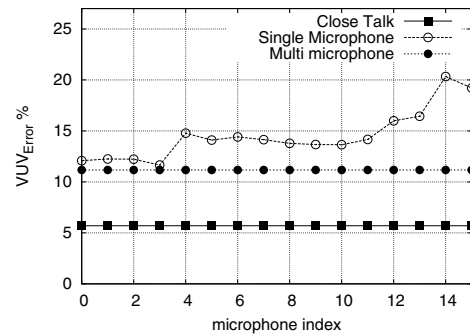


Figure 5: V/UV speech classification results with different input signals.

liability degree of each channel computed dynamically at each frame. MPF was used to estimate F0 and to extract two features that, in conjunction with an energy based feature, determine a feature vector used to classify voiced frames through a HMM based system. The effectiveness of using a DMN for the proposed voiced/unvoiced speech classifier was shown testing the system both on simulated data and on a real corpus. The results in which the DMN was exploited outperform always results obtained on a far microphone. The next activities will focus on the exploitation of the given techniques in distant-talking speaker identification and acoustic event classification for the interactive TV scenario under the DICIT project (<http://dicit.itc.it>).

8. References

- [1] F. Flego and M. Omologo, "Multi-microphone periodicity function for robust f0 estimation in real noisy and reverberant environments," in *Proc. ICSLP*, Pittsburgh, USA, 2006.
- [2] F. Flego and M. Omologo, "Robust f0 estimation based on a multichannel periodicity function for distant-talking speech," in *Eusipco*, Florence, Italy, 2006.
- [3] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE trans. SAP*, vol. 9, no. 7, pp. 727–730, Oct, 2001.
- [4] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, 2002.
- [5] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE trans. on acoustics, speech, and signal processing*, vol. ASSP-24, pp. 201–212, june 1976.
- [6] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE trans. on speech and audio processing*, vol. 7, pp. 333–338, may 1999.
- [7] S. Basu, "A linked-hmm model for robust voicing and speech detection," in *Proc. ICASSP*, Hong Kong, 2003, pp. 816–820.
- [8] S. Sagayama and S. Furui, "Pitch extraction using the lag window method," *Proc. of IECEJ*, 1978, (in Japanese).
- [9] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," in *JASA*, vol. 65, 1979, pp. 943–950.