A robust pitch extractor based on DTW lines and CASA with application in noisy speech recognition

Juan A. Morales-Cordovilla[1], Pablo Cabañas-Molero[2], Antonio M. Peinado[1], and Victoria Sánchez[1] ⋆

[1]Dept. of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain,
[2]Dept. of Engeniería de la Telecomunicación, Universidad de Jaén, Spain,
{jamc,amp,victoria}@ugr.es,cabanas@ujaen.es

Abstract. This paper proposes a robust pitch extractor with application in Automatic Speech Recognition and based on selecting pitch lines of a tonegram (a representation of the different pitch energies at each frame time). First, the tonegram and its maximum energy regions are extracted and a Dynamic Time Warping algorithm finds the most energetic trajectories or pitch lines from these regions. A second stage estimates the tonegram of the most energetic lines by applying Computational Auditory Scene Analysis rules which reject and group octave-related lines. The mean pitch of the speaker is estimated and the final pitch is estimated by rejecting lines which are outside from the mean pitch. The proposed pitch extractor is evaluated in a novel way - by means of the word accuracy of a Missing Data recognizer on Aurora-2 database.

Keywords: pitch extractor, pitch line, CASA, DTW, noise, robust speech recognition.

1 Introduction

Acoustic noise represents one of the major challenges for Automatic Speech Recognition (ASR) systems. Many different approaches have been proposed to deal with this problem [10, 1, 3] but if we consider voiced speech (i.e. not whispering speech) and the manner in which the auditory system works, pitch information can be a very useful cue to separate noise from speech and to obtain high performance in ASR [5, 7, 8].

One of the main challenges for pitch-based ASR techniques is that they need a robust pitch extractor. We can distinguish two stages in pitch extractors: a frame stage that obtains the pitch (or pitches) at each frame, and a post-processing stage which produces a final pitch decision. The result of the first stage is a representation indicating at each instant time, the energy or probability of observing

⋆ This work has been supported by the Spanish MEC/FEDER project TEC2010-18009 and partially funded by the DIRHA European project FP7-ICT-2011-7-288121.

the different pitch values. We will call *tonegram* to this representation and different tools such as difference-function [2], comb-filter [4] or auto-correlogram [5] can be employed to obtain it. The post-processing stage tries to estimate the final pitch by employing this tonegram and rules which help to distinguish the target pitch from possible noise pitches. The continuity and smoothness of pitch lines is the most common rule for speech signals as it is shown by the Hidden Markov Models (HMMs) or mode filters which many of the pitch extractors have [5, 8]. In addition, Computational Auditory Scene Analysis (CASA) rules, such as common limits (onset/offset) or even high level information [5], have been applied in order to group spectro-temporal pixels of the spectrogram and to obtain, as a result, a final pitch decision.

The goal of the paper is to show how the pitch lines can be extracted from a tonegram by means of a Dynamic Time Warping (DTW) approach, and how a final pitch decision can be obtained by means of a post-processing, inspired on CASA rules, of these lines. The advantage of working with pitch lines is that it let us associate to the lines different features (such as intensity, mean-pitch, space-localization, etc.) and later select the lines which fullfill the features of target speaker.

The structure of the paper is as follows. First, a block diagram gives an overview of the pitch extractor. Sec. 3 explains the proposed pitch extractor in greater detail. Sec. 4 presents the experimental framework and the Aurora-2 results by using a pitch-based Missing Data (MD) technique for ASR. The paper concludes with a summary and a discussion of future work.

2 System overview

The pitch extractor (Fig. 1) has a noisy signal of an utterance (the sum of clean speech and noise, $y = x + n$) as input. This signal is segmented and the autocorrelation of each frame is obtained to produce a tonegram. High energy regions of the tonegram are identified and their maximum energetic trajectories, obtained by means of a DTW approach, result in many pitch lines. We select a set of Maximum Energy Lines (M.E.L.) and their octave factors regarding their fundamental lines are estimated by using CASA rules. We relocate these lines at its fundamental period position, and estimate the tonegram which should be observed if only M.E.L. were presented with the addition of the corresponding octaves. We estimate the mean pitch of the speaker by means of this tonegram estimate and the final pitch *py* is obtained by discarding and selecting those lines which must correspond to the target speaker.

3 Pitch extractor

The most important blocks and functions of the proposed extractor are detailed below. Note that the parameters of the blocks were determined through preliminary experiments performed over a set of training sentences of Aurora-2 contaminated with noise.

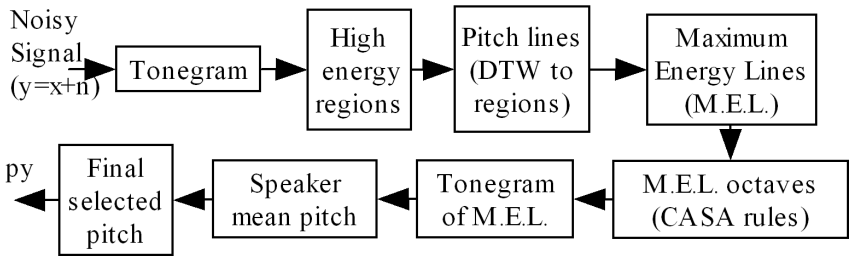

Fig. 1. Block diagram of the proposed pitch extractor.

3.1 Tonegram

In order to estimate the tonegram, the unbiased autocorrelation is employed due to the following properties: fast computation (by means of Fast Fourier transform), concentration of noise at first coefficients (if it is not correlated [8]), capacity of representing the pitch energy, and capacity to define regions when a tone is presented. The power tonegram at pitch value p and frame time t is:

$$TG_{pow}(p,t) = \frac{1}{FL-p} \sum_{i=p}^{FL-1} y_t(i) y_t(i-p) \tag{1}$$

where y_t $(i = 0, .., FL-1)$ is the noisy signal in frame t (length $FL = 256$, sampling frequency $8kHz$). The frame shift is $FS = 80$ samples and $p \in [pl, ph]$, where $pl = 10$ and $ph = 160$ samples define the range of human pitch. The power tonegram is passed through a square root function and normalized to $[0, 1]$ in order to obtain the final tonegram ($TG(p,t)$), which is a more suitable representation of pitch magnitude energy. Fig. 2 shows a tonegram from an Aurora-2 utterance.

3.2 High energy regions

The mean and the standard deviation of each temporal frame of the tonegram ($\mathbf{TG}(t)$) increase when a tone is presented, so we can estimate the instantaneous energy of the tonegram as follows:

$$\mathbf{E}_{TG}(t) = \mu_{\mathbf{TG}(t)} + \sigma_{\mathbf{TG}(t)} \tag{2}$$

where $\mu_{\mathbf{TG}(t)}$ and $\sigma_{\mathbf{TG}(t)}$ denote the mean and the standard deviation of a tonegram vector at time t. The instantaneous background energy $\mathbf{Eb}_{TG}(t)$ is obtained by passing $\mathbf{E}_{TG}(t)$ through a smoothing mean filter of length $WL/5$ samples (diameter $2 * WL/5 + 1$) followed by a minimum filter of length $WL/2$ samples. WL is 30 frames and refers to the expected mean Word Length. A tonegram pixel is classified with a boolean high energy indicator if $TG(p,t) > \mathbf{Eb}_{TG}(t)$. The high energy regions consist of connected high energy pixels. Regions with an area lower than $2 * WL/5$ pixels are deleted. Fig. 2 shows the resulting high energy regions. In the following, the l^{th} region will be denoted as $TG^l(p,t)$.

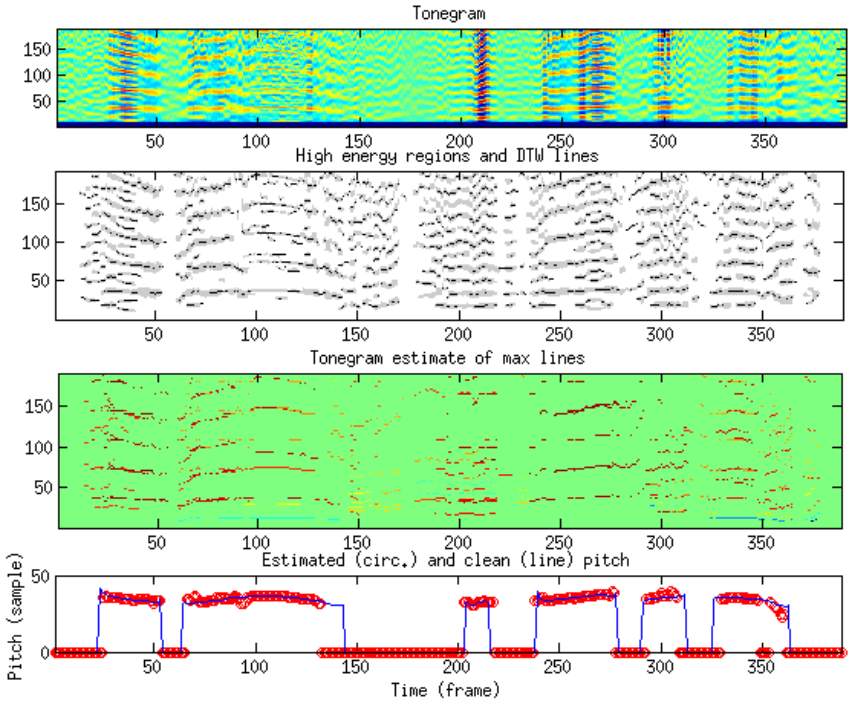

Fig. 2. Tonegram $TG(p,t)$ and its corresponding high energy regions with the DTW lines, tonegram estimate of Maximum Energy Lines (M.E.L) and pitch estimate for the FCJ_1396Z33A Aurora-2 utterance contaminated with babble noise at 0 dB.

3.3 Estimation of pitch lines based on DTW

Due to errors when estimating $\mathbf{Eb}_{TG}$, high energy regions can contain more than only one pitch line or even two or more crossed lines. An approach based on the maximum at each time in order to estimate the strongest energy line, can result in a discontinuous trajectory in these situations. Because of this, an approach based on searching for the path with maximum energy can be more suitable.

In other words, we can estimate a pitch trajectory through a region $TG^l(p,t)$ as the path that maximizes the global accumulated energy along axis t. For the sake of simplicity, in this section t will be a relative index ($t = 1, .., length^l$) where $length^l$ is the number of frames covered by the region. In order to find this path, we employ a method based on Dynamic Programming. The employed algorithm is quite similar to the well-known standard DTW technique [11], but introducing certain restrictions that we have found appropriate for pitch trajectory estimation.

Standard DTW is a pattern matching technique that has been used for decades in speech recognition, as well as in other areas, such as feature alignment in music [12]. Briefly, given a matrix $TG^l(p,t)$, DTW finds the warping path through the grid (p,t) that represents the "best" mapping between the

two axes according to $TG^l(p,t)$. This path is represented by a pair of warping vectors, $\mathbf{p}^l$ and $\mathbf{t}^l$, which give the coordinates of the path at every step i, that is, $\mathbf{p}^l = [p_1^l, p_2^l, \ldots, p_i^l, \ldots, p_I^l]$ and $\mathbf{t}^l = [t_1^l, t_2^l, \ldots, t_i^l, \ldots, t_I^l]$, where I is the number of steps in the path. In order to find the best path among all possible combinations, DTW minimizes the accumulated cost over the entire path. In our case, where $TG^l(p,t)$ represents energy (not cost), the optimal warping path can be defined as the one that maximizes the quantity $\sum_{i=1}^{I} TG^l(p_i^l, t_i^l)$, which measures the accumulated energy along the path.

In order to obtain a path that represents a meaningful pitch trajectory, some constraints must be imposed on the warping vectors. Firstly, the path must provide only a single pitch value for every frame, and secondly, the pitch trajectory must be smooth and continuous in frequency (and therefore, large hops in $\mathbf{p}^l$ should not be allowed). To satisfy both requirements, we impose the following local continuity constraints:

$$\mathbf{t}^l = [1, 2, \ldots, length^l] \tag{3}$$

$$\mathbf{p}^l = [p_1^l, \ldots, p_{length^l}^l] \quad \text{s.t.,} \quad |p_{i+1}^l - p_i^l| \leq h. \tag{4}$$

Clearly, the first constraint implies that each time frame will have only a single pitch, while the second one avoids pitch hops larger than h (in our experiments, we set $h = 3$ samples).

Taking into account these constraints, the DTW algorithm for finding the optimal trajectory through a region $TG^l(p,t)$ with size $P^l \times length^l$ can be summarized in two steps:

1. *Recursion*: For $1 \leq p \leq P^l$ and $2 \leq t \leq length^l$, compute

$$D(p,t) = \max_{p'}[D(p', t-1) + TG^l(p,t)], \tag{5}$$

whith initialization $D(p,1) = TG^l(p,1)$. Here, $D(p,t)$ can be interpreted as the maximum partial accumulated energy that can be obtained among all possible paths reaching the point (p,t). Observe that the maximization in (5) is performed only over the values p' from which (p,t) can be reached in a single step, in accordance with the constraint in (4). The best predecessor for each (p,t) is stored in ξ, i.e.,

$$\xi(p,t) = \arg\max_{p'}[D(p', t-1) + TG^l(p,t)]. \tag{6}$$

2. *Termination and Backtracking*: Finally, the optimal trajectory $\mathbf{p}^l$ is the path with higher global accumulated energy up to the end frame, yielding:

$$p_{length^l}^l = \arg\max_{p} D(p, length^l), \tag{7}$$

and the complete path is retrieved backwards as follows:

$$p_t^l = \xi(p_{t+1}^l, t+1), \quad \text{for } 1 \leq t \leq length^l - 1. \tag{8}$$

Fig. 2 shows the resulting DTW lines corresponding to each $TG^l(p,t)$ region.

3.4 Line features

Once the pitch lines have been extracted we must store the following data vectors of length $length^l$ for every line associated to the region $TG^l(p,t)$: $\mathbf{t}^l$, $\mathbf{p}^l$, $\mathbf{E}^l$ vectors with the time, pitch positions and instantaneous energy. We will also note E^l_{mean} the mean line energy, and t^l_{max}, t^l_{min}, p^l_{max}, p^l_{min} the corresponding maxima and minima.

3.5 Selection of Maximum Energy Lines (M.E.L.)

A vector with the line labels, corresponding to maximum mean energy (E^l_{mean}) at each time, is obtained and passed through a mode filter of length $WL/10$. This filter avoids including lines which are maximum for a very short time and its length is related to the temporal masking effect. The different filtered labels indicate the M.E.L. set. In the case of an energy tie, the line with lower pitch is selected because we are looking for the lines corresponding to the fundamental period. This situation will be addressed in Sec. 3.8.

3.6 Octave estimation

Any line corresponding to a fundamental pitch period should appear repeated at integer multiples, or horizontally in the tonegram. This can cause octave error when selecting M.E.L.s. The integer relation between the pitch of a maximum selected line lm and its fundamental line $lm0$ will be called the octave of lm ($o^{lm} = \mathbf{p}^{lm}/\mathbf{p}^{lm0}$) and is estimated by a grouping-line approach inspired on CASA [5] in these four steps:

1. Find horizontal lines close to lm: lines lh which fulfill this condition ($t^{lh}_{max} > t^{lm}_{min} \quad \& \quad t^{lh}_{min} < t^{lm}_{max}$) are selected.
2. Measure common movement, limit and intensity between lm and the horizontal lines lh as follows:

$$c^{lh}_{mov} = 1 - \frac{\sigma_{(\bar{\mathbf{p}}^{lm} - \bar{\mathbf{p}}^{lh}/f^{lh})}}{10} \tag{9}$$

$$c^{lh}_{lim} = 1 - \frac{|t^{lm}_{min} - t^{lh}_{min}| + |t^{lm}_{max} - t^{lh}_{max}|}{length^{lm}} \tag{10}$$

$$c^{lh}_{int} = 1 - \frac{|E^{lh} - E^{lm}|}{E^{lm}} \tag{11}$$

   where $\bar{\mathbf{p}}^{lm}$ and $\bar{\mathbf{p}}^{lh}$ indicate the common pitch part between lm and lh, and $f^{lh} = \mu_{(\bar{\mathbf{p}}^{lh}/\bar{\mathbf{p}}^{lm})}$ is the horizontal factor. Note that the maximum value for the common measures is always 1.
3. Select octave-related lines: the lines with common movement, limit and intensity bigger than $Th_o = (0.9, 0.9, 0.9)$ are the octave-related lines **lo** to lm. In case of not grouping lines, we try these other thresholds $Th_o = (0.7, 0.9, 0.9)$ and $Th_o = (0.9, 0.7, 0.9)$.

4. Estimate the octave of maximum line: If horizontal lines have not been selected, octave estimate is $\hat{o}^{lm} = 1$. If horizontal lines have been selected but not octave-related, $\hat{o}^{lm} = -1$. In other case, we estimate the octave considering that the f^{lh} of an octave-related line has to be an integer multiple of $1/o^{lm}$. For example, assuming $o^{lm} = 2$ the observed vector of octave lines should ideally be $\mathbf{f}^{lo} = 0.5, 1, 1.5, ...$ Taking this into account, the octave estimate is that which minimizes the distance between the observed and the ideal factor vector of an octave ($\hat{o}^{lm} = \arg\min_o(dist(\mathbf{f}^{lo}, \mathbf{f}_o^{ideal}))$). This distance is obtained by means of a clustering procedure and increases when the clustering error and the amount of not matched centroids (elements of $\mathbf{f}_o^{ideal}$) increases. The maximum possible tried octave is always $o_{max} = 6$.

3.7 Tonegram estimation of M.E.L.

The tonegram of M.E.L. is estimated as follows: we fill an empty tonegram with the original M.E.L. of Sec. 3.5 but relocated to their correct new position using the octave estimate ($\mathbf{p}_{new}^{lm} = \mathbf{p}_{orig}^{lm}/\hat{o}^{lm}$) and with the same original instantaneous energy. Also, the corresponding octave lines are put at integer multiples of $\mathbf{p}_{new}^{lm}$ and with the same energy. The lines with $\hat{o}^{lm} = -1$ are not moved but some possible octave lines are put at integer multiples and divisions of p_{orig}^{lm} and with the corresponding energy of the original tonegram. We do so because the octave is unknown. The maximum integer number, for adding octaves, is always limited to o_{max} in order to avoid the inclusion of too many lines. The features of this new tonegram are extracted and loaded in a structure as in Sec. 3.4. Fig. 2 shows this tonegram estimate.

3.8 Mean pitch estimation of the speaker

We select again the M.E.L. from the previous estimated tonegram in a similar way to Sec. 3.5 and a tonegram with these new M.E.L. is constructed. This tonegram will be denoted as TG_{perc} and can be considered as a representation of the perceived tones at each time if we are focusing our attention on maximum energy tones presented in the auditory scene. The total perceived energy of each tone ($E_{perc}(p)$) is obtained by summing neighboring channels separated one tone as follows:

$$\mathbf{E}_{perc}(p) = \sum_{t=1}^{nf} \sum_{\rho=[p*8/9]}^{[p*9/8]} TG_{perc}(\rho, t) \tag{12}$$

where nf is the number of frames and [] the round operator. Considering that, even at low SNRs, the majority of maximum tones correspond to the target speaker, we can say that the maximum of $\mathbf{E}_{perc}$ corresponds to the speaker mean pitch (p_{mean}).

3.9 Final pitch selection

If we suppose that the speaker pitch lines are concentrated around an interval of p_{mean} we can discard many lines from the M.E.L. tonegram of Sec. 3.7, so the l lines which do not fulfill this condition ($p^l_{max} > (2/3)p_{mean} \quad \& \quad p^l_{min} < (3/2)p_{mean}$) are deleted. In a similar way to Sec. 3.5, we select the M.E.L. of this deleted-tonegram and the corresponding pitches at each time of the line with maximum total energy conform our previous pitch estimate.

The previous unvoiced frames are those where pitch has not been detected. In the case that unvoiced frames are not detected, we suppose unvoiced the first and last 10 frames. In a similar way to Sec. 3.2 we obtain $\mu_{\mathbf{E}^u_{TG}}$ and $\sigma_{\mathbf{E}^u_{TG}}$ (the mean and the standard deviation of the instantaneous energy $\mathbf{E}_{TG}$ of unvoiced frames) in order to obtain an unvoiced background threshold. The instantaneous energy of the voiced frames ($\mathbf{E}^v$) is smoothed with a mean filter of length $WL/10$ samples and the frames with $\mathbf{E}^v < \mu_{\mathbf{E}^u_{TG}} + 5 * \sigma_{\mathbf{E}^u_{TG}}$ are labeled as unvoiced. Finally, the value of the previous pitch is made null at unvoiced frames and this is our final pitch estimate py. Fig. 2 also compares this pitch extraction with the clean pitch (extracted from the corresponding clean utterance).

4 Experimental framework and results

4.1 Experimental framework

The experiments reported here employ the Aurora-2 database which consists of digit utterances contaminated by different types of noises at different SNRs [9].

The evaluation of the pitch estimate will be done in a novel and useful way - by means of a pitch-based technique [6] for robust ASR. This technique has been presented in [7] and combines two complementary noises [a Voice Activity Detection noise (suitable for silence frames) and a tunnelling noise (suitable for voiced frames)] to estimate the noise spectrogram. This noise produces a soft Missing Data (MD) mask which is passed, together with the noisy spectrogram, to a marginalization MD recognizer. For the sake of simplicity, here, we will obtain a hard mask [3](instead of soft) which only requires the optimization of the threshold (and not also of the slope) to decide wether a feature is reliable or not. Clean train is always done and the HMM model features of the MD recognizer are the standards of Aurora-2 when the spectrogram is employed (9 Gauss/state, 23-LogMel-static+23-LogMel-delta feature vector, etc.. [7]).

4.2 Experimental results

Tab. 1 shows the different word accuracies achieved by different systems tested over the whole (set A, B and C) Aurora-2 database.

FE+CMN is the ETSI Front End (FE) with Cepstral Mean Normalization and acts on a classical cepstral recognizer [9]. The rest of the systems act on the MD recognizer explained above with different pitch extractors. *PEFAC* employs the pitch extractor proposed in [4] but, in order to improve its results, we apply

Table 1. Word accuracies obtained by different systems tested with Aurora-2 (set A, B and C) for different SNR values.

Systems	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean (20-0 dB)
FE+CMN	99.12	97.17	92.53	76.15	44.16	23.02	13.00	66.61
PEFAC pitch	98.67	93.56	84.69	69.29	55.23	37.30	18.31	68.01
Yin pitch	98.89	94.93	89.32	80.07	66.47	39.56	14.36	74.07
DTW-lines pitch (proposal)	98.20	95.07	90.14	80.93	66.15	39.06	14.90	74.27

the following post-processing: frames with voiced probability lower than 0.8 are selected as unvoiced. This decision is later passed though a mode filter of length 1 frame. Finally, we make null the pitch at unvoiced frames. *Yin* uses the extractor described in [2]. Frames with a normalized energy threshold lower than 0.8 and gross aperiodicity bigger than 0.95 are considered unvoiced. *DTW-Lines* employs the proposed pitch extractor. The optimum threshold of the masks was $-3dB$ in all cases, except for the *PEFAC* approach ($0dB$).

We can see that our pitch extractor outperforms all the extractors on average. In clean conditions, our pitch extractor does not obtain as good results as the others probably because the background energy thresholds of Sec. 3.2 and 3.9 avoid the detection of some weak regions and pitch values respectively.

5 Conclusions

This paper has proposed a pitch extractor for ASR based on the assumption that the most energetic pitch lines of the tonegram, around a speaker mean pitch estimate, correspond to the speaker pitch. The pitch lines have been extracted with a DTW approach and CASA rules have been employed to group and reject lines. The proposal has been evaluated on a robust ASR system showing high performance. Regarding future work, the results at clean and noisy conditions could be improved by means of a better estimation of the background energy threshold and a better application of CASA rules in the selection of the target speaker lines. Also we would like to test this scheme on another more robust tonegram (such as the difference function [2]), and on the two-talker recognition problem [5] by using the line features (such as the intensity, mean-pitch or even space-localization) together with high level information (provided by Speech Fragment Decoding [1, 5]) in order to separate the pitch lines of the two speakers.

References

1. J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005.
2. Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111 (4):1917–1930, 2002.

3. M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, 2001.
4. S. Gonzalez and M. Brookes. A pitch estimation filter robust to high levels of noise (pefac). In *EUSIPCO*, 2011.
5. N. Ma, P. Green, J. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Communication*, 49:874–891, 2007.
6. Juan A. Morales-Cordovilla. *Pitch-based technique for robust speech recognition.* PhD thesis, Dept. of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain, 2011.
7. Juan A. Morales-Cordovilla, Ning Ma, Victoria Sánchez, José L. Carmona, Antonio M. Peinado, and Jon Barker. A pitch based noise estimation technique for robust speech recognition with missing data. In *ICASSP*, pages 4808–4811, May, 22-27 2011.
8. Juan A. Morales-Cordovilla, Antonio M. Peinado, Victoria Sánchez, and José A. Gonzalez. Feature extraction based on pitch-synchronous averaging for robust speech recognition. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 19(3):640–651, March 2011.
9. D. Pearce and H. G. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP*, volume 4, pages 29–32, 2000.
10. Antonio M. Peinado and Jose C. Segura. *Speech Recognition over Digital Channels.* Wiley, 2006.
11. Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition.* Prentice-Hall, 1993.
12. R. J. Turetsky and D. P. Ellis. Ground-truth transcriptions of real music from force-aligned midi syntheses. In *Int. Conf. Music Inf. Retrieval (ISMIR)*, pages 135–141, 2003.