



A Multipitch Tracker for Monaural Speech Segmentation

André Coy and Jon Barker

Department of Computer Science, University of Sheffield
 211 Portobello Street, Sheffield, S1 4DP, United Kingdom
 {a.coy, j.barker}@dcs.shef.ac.uk

Abstract

This paper presents a novel algorithm for forming coherent harmonic fragments from a mixture of speech sources. A multiple pitch detection algorithm is used to produce pitch candidates which are tracked using a pair of parallel HMMs. One novel aspect of the technique is that it systematically models pitch doubling and halving errors, thereby facilitating the identification of smooth pitch segments even in the absence of the fundamental frequency. The system does not face the problem of incorrect source assignment that can occur when sources have similar fundamental frequency or are harmonically related. An evaluation of the technique shows that the algorithm's emphasis on tracking coherent segments leads to the formation of speech fragments with high coherence, indicating a more reliable segmentation of the harmonic speech regions.

Index Terms: multiple pitch detection, source separation

1. Introduction

The challenge addressed is the detection of harmonic regions in sounds combined over a single communication channel. Of particular concern are mixtures where one or more source is speech. The goal is usually to isolate regions dominated by a single source for use in re-synthesis or automatic speech recognition. This work is concerned with segmenting speech into fragments (spectro-temporal regions that can be assigned to a single source) to be used for recognition of the sources in the mixture. The framework is such that an arbitrary number of sources can be segregated, however the current study focuses on forming fragments from speech mixed with speech from a single talker.

It has long been recognised that both top-down and bottom-up processes must interact in order to accurately group fragments of speech and assign them to a single source [1]. Rather than using harmonicity as the sole criteria for segment grouping (bottom-up grouping), the grouping process should be informed by a top-down technique that takes into account the information available from a rich spectral representation of speech.

The multipitch tracker presented uses a novel Hidden Markov Model (HMM) based pitch tracking algorithm to perform bottom-up decomposition of mixed speech into harmonic segments thought to belong to a single source. These segments (regions of continuous voicing divided by unvoiced regions) are transformed into coherent spectro-temporal fragments - a coherent fragment is one completely dominated by a single source. The coherence of speech fragments is extremely important - it is better to produce a completely coherent fragment which can be directly assigned to a single source than to produce an incoherent one that cannot be correctly assigned to any source.

The top-down grouping can be performed by the Speech Fragment Decoder (SFD), which, accepts coherent fragments and makes simultaneous grouping and word-sequence decisions using information contained in HMM models of speech [2]. In the past, fragments were developed using processes that tracked pitch across the entire utterance emphasising the formation of complete pitch tracks for each source [3]. While the current algorithm also forms complete pitch tracks it goes further in attempting to maximise the coherence of each voiced segment.

2. The Multipitch Tracker

A schematic of the system is shown in Figure 1. The focus of this paper is the multipitch tracker (MPT) which uses computational models of primitive auditory scene analysis to extract voiced speech regions.

2.1. Pitch Detection

An autocorrelogram (ACG) based multiple pitch detection algorithm (MPDA) is used to detect all the relevant pitch periods present in the signal. The sampled mixture is passed through a 64 channel gamma tone filter bank spaced equally on an equivalent rectangular bandwidth (ERB) rate scale with centre frequencies between 50 and 8000 Hz. The signal is then framed using a 35 ms window with a 10 ms frame shift. The filter output is used for direct computation of the ACG while the envelope of the filter response is used to compute the envelope ACG (eACG). The envelope response is computed because of the generally unresolved harmonics of high frequency channels which tend to be amplitude modulated with a beating response corresponding to the fundamental frequency of the signal. A normalised summary ACG (sACG) is then computed from the low frequency channels of the ACG. The peaks from the sACG that exceed an empirically derived threshold θ_s are stored. Up to four peaks are stored in an effort to minimise information loss from the signal.

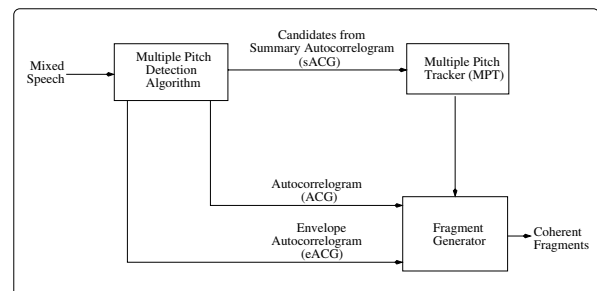


Figure 1: Schematic of the fragment generation system.

This work was funded by EPSRC grant GR/R47400/01

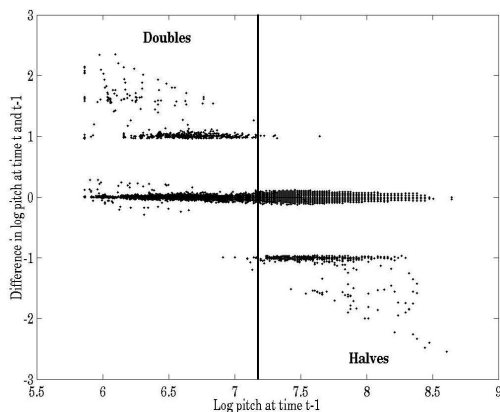


Figure 2: Illustration of the threshold for pitch doubling and halving in female speech. The data to the left of the line represent points where the pitch at time t is at least twice the pitch at time $t - 1$ (doubles). To the right of the line the pitch at time t is at most half the pitch at time $t - 1$.

2.2. Modelling Pitch Dynamics

The pitch candidates output by the sACG are assumed to be generated by a number of Markov processes separated into those which generate the correct pitch tracks and those that output a number of distractor observations. The full system requires the modelling of both types of process.

Each pitch track is generated by a two state HMM with transition probabilities $p(v | v)$, $p(v | u)$, $p(u | v)$, $p(u | u)$, where v and u represent the harmonic and inharmonic condition, respectively. The tracker works in the log domain, thus an observation, f_t , at time t is transformed to $f'_t = \log_2(f_t)$. In the harmonic state, an observation at time t is drawn from $p(f'_t | f'_{t-1})$ if the previous observation is harmonic or $p(f'_t)$ if it is inharmonic¹. In order to capture pitch segments from N sources simultaneously, N models are run in parallel.

The distractor points are modelled by an independent noise model - $p(d_n)$ - where each point is assumed to be an identical, independently distributed (iid) random variable. The noise model generates all the observations not generated by the pitch track models. The number of distractor points in each frame follows a distribution, $p(N_d)$ represented as a histogram of probabilities. Gender dependent distributions of pitch dynamics are estimated from clean speech by analysing the pitch of the utterances in the AURORA 2 training set [4]. These pitches are calculated by running the above-mentioned pitch finding algorithm and choosing the two highest peaks in the sACG as the pitches of the sources. Drawing an observation from $p(f'_t | f'_{t-1})$ is statistically equivalent to adding a difference, $D = f'_t - f'_{t-1}$ to f'_{t-1} , with D being drawn from the distribution $p(D | f'_{t-1})$. The difference D , is the per-frame pitch change and is expressed as the ratio of neighbouring log pitch estimates $\log_2(\frac{f_t}{f_{t-1}})$. The ratio will remain within a band of values close to 0, except where there is pitch doubling or halving; where the values will be closer to 1 (doubles) and -1 (halves). A plot of D against $f'(t - 1)$ (Figure 2) shows that pitch doubling is more likely to occur at values below dp_t - approximately $\log_2(148Hz)$ - while halving occurs at higher values. The generative model can be

¹Here $p(f'_t)$ represents the prior distribution of pitch values in the data set while $p(f'_t | f'_{t-1})$ captures the per-frame pitch dynamics.

approximated by two separate distributions: $p_{double}(f'_{t-1})$ for values of f'_{t-1} less than dp_t and $p_{half}(f'_{t-1})$ for higher log pitch values. These distributions are modelled by Gaussian mixture models (GMM) estimated from the training data. The centres of the components are set to -1 , 0 and 1 , while the variances and weights are estimated using maximum likelihood. The prior distribution of pitch values is also estimated using a mixture of Gaussians. HMM state transition probabilities are calculated for male and female speech using the log transformed data. The transition probabilities capture the dynamics of voicing; how likely is it that voicing will begin, continue or end.

For noise model estimation, the peaks of the sACG (up to four peaks) with values above θ_s are chosen. The value of θ_s acts as a threshold for source voicing. Peaks above this value indicate output from a voiced source. For each frame, the peaks that correspond to the *a priori* pitch of the two sources are removed; the remaining peaks are used to build the noise model distribution. *A priori* pitch estimates are obtained for each source by tracking the pitch of the unmixed signals using Snack (an open source version of ESPS/waves+) [5]. The pitch represented by a peak from the sACG is considered to match the *a priori* pitch if it lies between $\pm 5\%$ of the *a priori* pitch. This accounts for the potential variation brought about by mixing the sources. A Gaussian mixture model is used to estimate the distribution of distractor points. The number of noise points (peaks that remain after the removal of the pitch peaks) is calculated for each frame and the probability of different numbers of noise points per frame, $p(N_d)$, is calculated.

2.3. Tracking Segments

The algorithm is implemented to track the pitches of a mixture of two speakers' speech. For this condition, peaks with values above θ_s (set to 0.8) are chosen from the sACG and used in the pitch tracking algorithm. A maximum of four peaks are chosen for each frame. Within a frame, there are either two, one or no active pitch sources; similarly, the number of distractor points may vary. All the possible interpretations of a set of observations is considered with the exception that no single candidate is assigned to both sources simultaneously. Considering 4 peaks the possible interpretations are: both speakers are in the unvoiced state and all observations are due to the noise model (1 interpretation); only speaker 1 is voiced and accounts for 1 of the 4 observations (4 interpretations); only speaker 2 is voiced accounting for 1 observation (4 interpretations); both speakers are voiced and 2 peaks are noise (4×3 interpretations). Where a source is hypothesised to have entered the unvoiced state, the inclusion of the noise model keeps the number of terms in the probability calculation constant. For two neighbouring frames, all combinations of candidates are considered and a score is generated for each combination. A lattice of hypotheses is formed and the best path through the pitch space is calculated using the Viterbi algorithm. The path describes the most likely interpretation of the observations, assigning observations to models and indicating the state of voicing for each speech region. Each pitch track is a sequence of continuously voiced segments divided by unvoiced regions. Figure 3 shows a typical result of the outlined process. Each segment is deemed to be coherent (i.e. belong to a single source).

2.4. Fragment Generation

Each pitch segment generated by the MPT is associated with a fragment of speech; the aim of the fragment creation process is

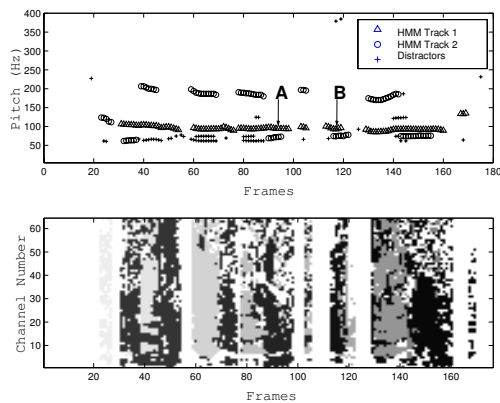
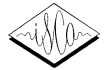


Figure 3: The pitch segments (top) and fragments formed from them (bottom) are shown for a mixture of male and female speech. The distractor points that the noise model has produced are overlaid. The segments marked **A** and **B** show where the sub-harmonic is tracked, continuing the segment, in the absence of the fundamental due to the inclusion of a model of pitch doubling and halving. Notice also that overlapping segments produce separate fragments even when the sub-harmonic is tracked.

to produce a set of coherent fragments (fragments completely dominated by a single source). The fragment is constructed in a frame by frame process where each channel of the ACG and eACG are examined for peaks that match to either (or both) of the pitch candidates output by the MPT. If the peaks detected in a channel match candidates from the MPT, then that channel is considered to be reliable for that frame. Figure 3 shows the fragments derived from each segment of voiced speech - each fragment is a different shade of grey. When there are two sources active in a frame the source with the highest peak is chosen as the dominant one. The decision however, has to be softened because the dominance of a source within that channel is not assured. Rather than expressing total confidence in the dominance of the chosen source, a soft decision is made by considering the relative contributions of both sources in the channel (for details see [6]). For the purpose of fragment decoding a soft mask is built up using the soft values for each spectro-temporal point. This is useful for fragment decoding where the recogniser applies the soft value as a weight on the dominance of a given source i.e., to what extent it is masked. In the soft mask, each spectro-temporal point is assigned a value between 0.5 and 1 where completely dominant points are assigned a ‘1’. A value of ‘0.5’ indicates that both sources were equally energetic and the decoder has to choose which is more likely to have generated the fragment.

3. Evaluation

3.1. Fragment Size and Coherence

The pitch segments generated from the MPT are used to generate fragments which are evaluated for their coherence. The coherence of a fragment is calculated by comparing it to the *a priori* mask of each unmixed source. Each fragment usually matches one source more than the other. If the fragment fits within the boundaries of a single source without overlap, it is 100% coherent. Otherwise its coherence is calculated by: $100 \times \sum w_i / (\sum w'_i + \sum w_i)$, where the w_i are soft values of the

fragment which fall within the boundaries of the mask it best fits and the w'_i are the soft values which overlap the other mask.

Experiments are performed using simultaneous speaker data constructed from the Grid corpus [7]. The Grid corpus consists of utterances spoken by 34 speakers reading sentences of the form: <command><colour><preposition><letter><number><adverb> e.g., “place white at L 3 now”. In the present study pairs of end-pointed utterances have been artificially added at a range of target-masker ratios (TMR). The test set has 600 utterance pairs at each TMR: 200 same speaker, 200 same gender (but different speakers), and 200 mixed gender.

The effect of segment grouping on fragment coherence is investigated using 100 utterances randomly selected from the test set. The voiced segments of each utterance are treated as isolated ‘pitch tracklets’ and three separate grouping mechanisms are applied creating three separate sets of fragments. For the first set *no post-processing* (**NP**) is applied to the segments; a new fragment is formed whenever there is a break in voicing. For the second set (**CL**) a simple *clustering* algorithm groups each voiced segment creating two continuous pitch tracks from which two fragments are formed, one for each source. The third set is formed by *partial grouping* (**PG**), where neighbouring tracklets are grouped if their absolute pitch difference lies within a distribution of absolute pitch differences estimated from training data. The coherence of fragments formed from all three techniques is compared. The coherence values presented are an average for all fragments in the utterance.

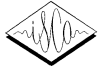
	0 dB		6 dB	
	Frag. Size	Coherence	Frag. Size	Coherence
CL	3380.7	61.5	3312.1	61.7
PG	377.3	74.9	369.0	74.5
NP	317.3	77.4	308.6	77.6

Table 1: A comparison of average fragment coherence and fragment for three algorithms at different TMRs. **NP** is the system presented, which does no post-processing on the segments. **CL** uses clustering to produce two segments while **PG** groups neighbouring segments based on f_0 difference

Table 1 reveals an inverse relationship between coherence and fragment size. Smaller fragments are (in general) more coherent than larger ones; this relationship holds across different TMRs, indicating that sequential grouping constraints which employ only f_0 dynamics can have a detrimental effect on fragment coherence. The process which grouped segments across the entire utterance (**CL**) produced the least coherent fragments showing that assigning segments to sources without top-down knowledge can produce fragments in which sources significantly overlap. Even minimal grouping (**PG**) can negatively affect coherence. It is worth noting that the fragments retain their coherence even as TMR decreases from 6dB to 0dB.

3.2. Fragment Coherence and Speech Recognition

Although the fragments generated by the MPT are useful for speech recognition, speech fragment decoding is not the focus of this paper (for a full set of recognition experiments on the Grid corpus using fragments generated by the MPT see [8]). However, the relationship between coherence and recognition scores can be examined by comparing the scores achievable



when fragments are correctly grouped with those obtained by decoding a set of completely coherent fragments. The grouping is done using *a priori* masks of the unmixed sources to group the subset of fragments that best match each source. Three types of fragments are presented: i) *completely coherent* fragments (CC in figure 4), where all regions belong to the target source, ii) fragments formed from the *complete tracks* (CT) output by the MPT - one fragment per source and iii) fragments formed from the continuous *voiced segments* (VS) - potentially several fragments per source. Completely coherent fragments are generated by removing the regions of each grouped subset of iii) that are outside the boundaries of the mask it best matches.

The utterances in the Grid corpus test set are mixed such that the ‘colour’ for the target utterance is always ‘white’; the masking utterances never contain the ‘colour’ ‘white’. The task is to recognise the letter and digit spoken by the target speaker (i.e. by the person who says ‘white’). Speaker dependent word-level HMMs were trained for each of the 34 speakers in the corpus. The words were modelled with a straight-through, no skip topology using 12 states each and 7 diagonal covariance Gaussians. Missing data decoding was performed on the entire test set using the grouped fragments and a grammar reflecting that the target always spoke the colour ‘white’.

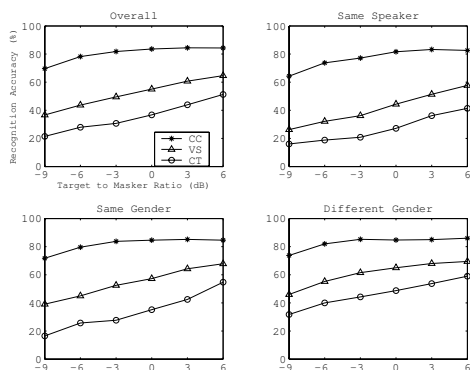


Figure 4: A three way comparison of speech recognition scores for fragments formed from complete bottom-up pitch tracks (CT) with fragments formed from voiced segments (VS) and completely coherent fragments (CC).

Figure 4 clearly shows that coherence directly affects recognition scores. The difference in scores for the conditions is due solely to coherence as all three were decoded under the same conditions. As suggested by section 3.1 the smaller fragments formed from continuous voiced segments (VS) are more coherent than fragments formed from the complete tracks (CT). The more coherent fragments yield higher accuracy. This underscores the importance of ensuring the coherence of segmented speech. The results also show the potential for detecting speech segments in mixtures of same gender sources, as even the most difficult case (Same Speaker) shows reasonable accuracy for the fragments formed from the segmented tracks.

4. Discussion

The MPT is shown to produce highly robust pitch segments leading to the production of coherent fragments. It has also been shown that greater coherence leads to better recognition results. The algorithm presented performs well even when a multiple or sub multiple of the fundamental is detected as these are neither corrected nor ignored, but incorporated into the segment where

appropriate. This is done for two reasons. Firstly to avoid errors of incorrect segment assignment or the complete omission of segments. The second reason is linked to the end use of the segments. For speech recognition, if a fragment of speech is extracted from a mixture based on the harmonics of a source and not the fundamental, the same information will be made available to the decoder. Other multipitch algorithms (e.g., [9]) may not be suitable for fragment generation because of potential source assignment errors that can occur when sources are harmonically related or when the fundamental is missing (below the voicing threshold or completely absent). By emphasising the coherence of each segment, the algorithm presented lessens the probability of such errors occurring.

There are however, several issues that require closer attention. Firstly the MPT’s tuning parameters: the voicing threshold θ_v , and the halving/doubling threshold dp_t . Experiments have shown that dp_t can be varied within a small range of values without significantly affecting performance. The voicing threshold determines which regions of speech are harmonic. If this is set too high, smoother segments may be formed but less information is retrieved from the sACG; if set too low, and more distractor points will emerge. If a source is strongly harmonic at least one of its harmonics will be present in the summary. By retaining several peaks the current method attempts to detect it.

The assumption that pitch segments are generated by a first order Markov process produces good results, however the framework can readily be extended to model the pitch trajectory using a second order Markov model. Further, using Gaussians to model the pitch dynamics leads to an imperfect fit. However, as a first approximation it provides reasonable results. The assumption that the distractor points are iid noise does not account for the relationship between the distractors and the pitch. There is scope for addressing this in the form of a re-estimation of the model parameters. Whilst this will be a further approximation, it may serve to improve the models in a systematic way.

5. References

- [1] C.J. Darwin and C.E. Bethell-Fox, “Pitch continuity and speech source attribution,” *Journal of Experimental Psychology: Human Perspective and Performance*, vol. 3, no. 4, pp. 665–672, 1977.
- [2] J. Barker, M. Cooke, and D. Ellis, “Decoding speech in the presence of other sources,” *Speech Communication*, vol. 45, pp. 5–25, 2005.
- [3] A. Coy and J. Barker, “Recognising speech in the presence of a competing speaker using a ‘speech fragment decoder’,” in *Proc. ICASSP ’05*, 2005, vol. 1, pp. 425–428.
- [4] H.G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. ICSLP ’00*, 2000, vol. 4, pp. 29–32.
- [5] K. Sjolander, “The snack sound toolkit version 2.2b1,” <http://www.speech.kth.se/snack/>, 2002.
- [6] A. Coy and J. Barker, “Soft harmonic masks for recognising speech in the presence of a competing speaker,” in *Proc. Interspeech ’05*, 2005, vol. 1, pp. 2641–2644.
- [7] M.P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *Journal of the Acoustical Society of America*, submitted.
- [8] J. Barker and A. Coy, “Recent advances in speech fragment decoding techniques,” in *Proc. ICSLP 2006*, accepted.
- [9] M. Wu, D.L. Wang, and G.J. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Transactions on Speech and Audio Signal Processing*, vol. 11, pp. 229–241, 2003.