Robust Multi-Pitch Tracking: a trained classifier based approach

Michael Kellman[§] and Nelson Morgan[*§]

TR-16-002

August 2016

Abstract

Pitch determination algorithm (PDA) performance typically degrades in the presence of interfering speakers and other periodic sources. We propose a multi-pitch determination algorithm (MPDA) that will detect and estimate two pitch tracks, a dominant and an interferer. Our method aims at being robust to various levels of combination of two speakers. Similar to the Subband Autocorrelation Classification (SAcC) method, we present a classifier based approach trained on compressed correlogram features. In contrast we train our classifier to detect all periodic sources, allowing for multiple speakers to be present. Viterbi decoding over a Markov chain of possible pitch and multiple speaker states is used to generate significant and continuous pitch tracks. We will compare our proposed algorithm against another MPDA and evaluate the performance of the methods with metrics extended from traditional PDAs.

§ University of California, Berkeley, Berkeley, California, 94720
* International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, California, 94704

This work was partially supported by funding provided to ICSI through a National Science Foundation grant IIS-1320260. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation.

Introduction

Speech's pitch track information has a variety of applications in the areas of speech compression and speaker identification. Knowledge of multiple pitch tracks present in a single channel signal can aid in the speech separation problem. Most Computational Auditory Scene Analysis (CASA) systems exploit the information of pitch and periodicity to perform speech separation and recognition (Wang et al, 2006). The performance of these CASA systems are tightly tied to the reliability of their input features, motivating the research of reliable and robust methods for extraction of pitch tracks.

In the presence of interfering speech sources the performance of traditional single pitch tracking algorithms is greatly degraded. We present a method for the tracking the pitch of both the dominant and interfering source. In the situation where the interfering speaker's contribution to the mixture is insignificant, determination of the dominant speaker's pitch track simplifies to the single pitch tracking problem. When the dominant and interfering speaker contribute equally to the mixture, the concept of having a single dominant speaker is ill-defined and it is much more difficult to extract either pitch track. We are interested in this degenerate case as it is a difficult case for pitch tracking algorithms.

Related Works

Autocorrelation-based pitch tracking methods have shown much success. Wu et al. (2003) have developed a robust MPDA that is evaluated on clean and noisy speech (Wu algorithm). The Wu algorithm combines pitch peak information from the subband autocorrelation domain to form an estimate of the pitch posterior distribution at a discrete time step. They then perform Viterbi decoding over a Hidden Markov model (HMM) to generate continuous pitch tracks.

Lee et al. (2012) present a PDA, Subband Autocorrelation Classification (SAcC), that uses a compressed subband autocorrelation representation to train a multi-layer perceptron (MLP) that estimates the pitch state posteriors. Viterbi decoding is then performed over the pitch state space to generate continuous pitch tracks.

Both of the related work's methods and the proposed method follow a similar sequence of computational blocks to achieve their goals, listed as follows: a feature extraction component, a pitch state estimation component, followed by a Viterbi decoding component. All three methods use a similar feature representation referred to as the subband autocorrelation. The pitch state estimation component's goal is to estimate the likelihood of a pitch being present from the observed speech's features at each time point. Finally, the decoding component's goal is to interpret

the distributions across time points to form continuous pitch tracks. These components will be discussed in the following sections.

In the analysis of the Wu algorithm they report solely on their ability to estimate the dominant pitch track and disregard the performance of the interfering pitch track. In the degenerate case, when two speakers contribute equally to the mixture, the performance of the algorithm will be degraded due to the ambiguity of which speaker is dominant. Our proposed method aims to build on this by reliably extracting two pitch tracks present in the speech mixtures and will be evaluated based off of the integrity of both. We aim to achieve this goal by expanding and improving upon the SAcC method, due to the high quality of its performance in single pitch tracking setting. Our modifications to the SAcC method are highlighted in the second and third outlined components: a restructuring of the neural network outputs to allow for an intuitive labelling of all of the pitch tracks present and the development of a multiple speaker Markov model to allow for the tracking of multiple pitch tracks.

Subband Autocorrelation Features

For both of these methods the input speech, $x[n]$, goes through a similar pipeline to extract subband autocorrelation features. Speech is decomposed into a set frequency subbands, $x_l[n]$ for $l = 1\ to\ S$, using a cochlear filter bank. The traditional model for a cochlear filter bank is a set of fourth-order gammatone filters with center frequencies uniformly distributed from 80 Hz to 5kHz and bandwidths set according to the Equivalent Rectangular Bandwidth scale. At this point the $x_l[n]$ is discretized into 10ms chunks and a N-point normalized autocorrelation is computed (n is indexing time, k indexing in correlation lag, and l is indexing the subband).

$$A(l,n,k) = \frac{\sum_{p=-N/2}^{N/2} x_l[n+p]x_l[n+p+k]}{\sum_{q=-N/2}^{N/2} x_l[n+q]x_l[n+q] \sum_{r=-N/2}^{N/2} x_l[n+k+r]x_l[n+k+r]}$$

Pitch State Estimation and Decoding

The Wu algorithm then performs peak detection within each of these subbands to localize periodic energy. The resulting pitch peak energy is spread along the periodicity dimension according to an empirical Laplacian fit and summed across the frequency subband dimension. They interpret the merged results as an estimate of the probability distribution over possible pitches. Viterbi decoding is then performed over an HMM to enforce sequential consistency to obtain continuous pitch tracks.

The SAcC algorithm generates probability distributions from the output of a MLP classifier. To make the training of the classifier feasible the raw subband autocorrelation features are reduced in dimension by representing each subband at each time point as its top K principal components. This reduces the input feature

dimensionality from S by N to S by K. The ground-truth pitch data from the speech corpus is used to label the training data, more on this later. The pitch state space is discretized into 67 logarithmically spaced frequency bins with the addition of a `too high', `too low', and `unvoiced' bins. The ground-truth pitch is mapped to its closest quantized bin for each time frame. These 70 bins are used as the output units of the MLP. Similar Viterbi decoding is performed to form continuous pitch tracks.

Methods

Below we will discuss how our MLP classifier is structured and trained, as well as how the multiple speaker Markov model is generated and used to decode the posterior pitch states.

MLP Classifier

Our method classifies the reduced dimension subband autocorrelation features into all the pitch states that apply. This allows for multiple pitch labels to be assigned to a particular time frame. The architecture consists of a single hidden layer with 1000 hidden units. There are S by K input units, representing the dimensionality of our input features, and 70 output units representing the number of states in our quantized pitch space.

We train our MLP on the features extracted from mixed speaker speech. Our training data is labeled so that the MLP classifies all periodicities present in each time frame. When neither of the speakers is voiced then the 'unvoiced' bin is labeled. When one speaker is voiced, that respective speakers' pitch bin is labeled. When both speakers are voiced then both speakers' pitch bins are labeled. The label vectors are normalized at every time frame, so that the MLP estimates valid posterior distributions. Figure 1 is the output of the trained MLP on a mixed speaker speech.

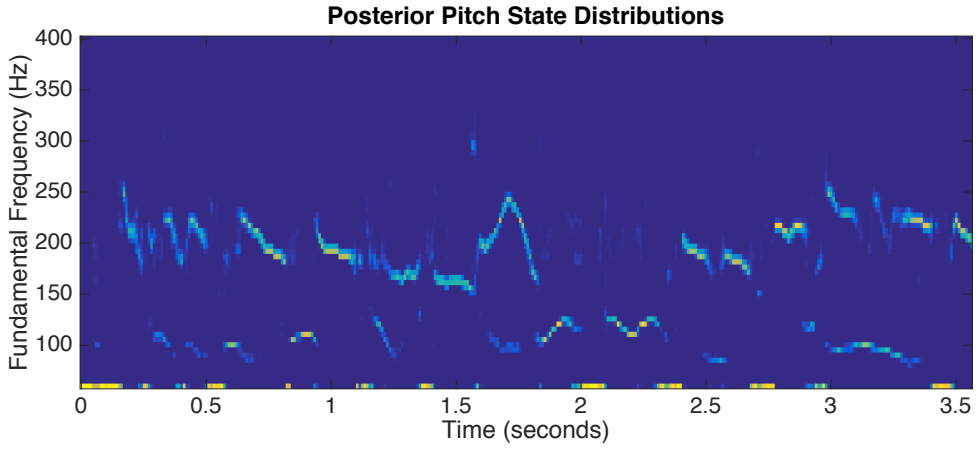

Figure 1: Resultant MLP posterior pitch state distributions for 356 consecutive 10 ms time frames on a mixture of 2 speakers.

Multi Speaker Markov Model

The MLP classifier estimates the posterior pitch state distribution at each discrete time step. We then divide our resulting distributions at each time frame by the pitch state prior to extract the likelihood of the observation given the current pitch state. We empirically estimate the prior of unvoiced and voiced frames and use a uniform prior on all voiced pitch states. These distributions can be decoded to give final continuous pitch tracks.

We have constructed a Markov chain that allows for multiple fundamental frequencies to be tracked. The Markov chain is constructed from the cross of two identical Markov chains that each span the single speaker pitch space. These single speaker Markov chains are constructed in a similar manner as in Lee and Ellis. The resulting state space allows for two speaker's pitch tracks to be simultaneously decoded and to form a maximum likelihood estimate of a pair of continuous pitch tracks. The top pane of Figure 2 visualizes the results of our decoding.

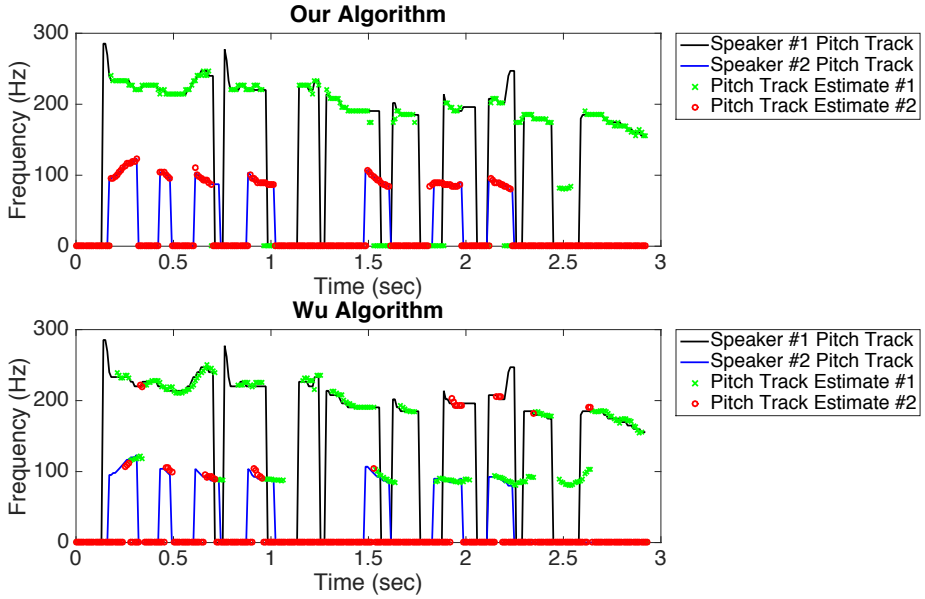

Figure 2: Comparison of results for artificially mixed TIMIT utterances. Our method's pitch track estimates of monaural mixed speaker speech plotted against ground-truth labels (top pane). Wu et al. method's pitch track estimates of monaural mixed speaker speech plotted against ground-truth labels (bottom pane).

Error Metric

PDA evaluation metrics have been established in previous publication by Rabiner et al. (1976). These have been improved upon in Lee et al. (2012) and expanded upon to evaluate MPDAs in Wu et al. (2003). Defined below are several metrics that quantify the ability of a MPDA to correctly determine the pitch tracks of multiple

speakers. These metrics can be divided into two categories: speaker error and pitch tracking error. Speaker error accounts for the improper number speaker pitch tracks in a particular frame. For example $E_{0\to1}$ is the percentage of unvoiced frames being misclassified as single speaker voiced frames. This is expanded on for all plausible cases of speaker error, with the notation of $E_{n\to m}$, the percentage of frames with n number of speakers being misclassified as a frame with m number of speakers.

Pitch tracking error accounts for error in the value of pitch estimated and this is evaluated on the frames that have the correctly classified number of speakers. This is further divided into two categories of Fine Error (FE) and gross error (GE). GE is the percentage of pitch estimates that deviate by more than 20% from their ground-truth label. FE is the standard deviation of pitch estimates from the ground-truth labels. These two metrics are evaluated separately in the single and multiple speaker cases. The single speaker case occurring when one of speakers is temporarily silent and the multiple speaker case occurring when neither speaker is silent.

Experiments

The KEELE (Plante et al, 1995) and FDA (Bagshaw et al, 1993) corpora are used for training purposes. Both provide laryngeal frequency contours that we extract ground-truth pitch labels from. The KEELE corpus contains the speech of 10 speakers each with 30 seconds of pitch-labeled data and the FDA contains 2 speakers each with 50 three second chunks of pitch-labeled data. We generate our training and cross validation data from the FDA and KEELE corpora by artificially mixing chunks of audio files to form 2000 two speaker utterances, each roughly 5 seconds in duration. Each speaker contributes equally in each training mixture, so that neither speaker is dominant.

Testing data is derived in a similar fashion from the TIMIT corpus (Garofolo et al, 1993). We extracted ground-truth pitch labels from our single speaker data prior to mixing with the SAcC algorithm. Lee et al. (2012) reported the SAcC algorithm performed highly in the presence of noise and was in agreement with other PDAs and therefore should act suitably as our ground truth. We generated 20 two speaker utterances each roughly 5 seconds in duration and ran several experiments varying the weight of each speaker's speech in the mixture. We ran experiments with equal contribution (0 dB), 5 dB, 10 dB, and 15 dB contribution. We evaluated our method side by side with the Wu algorithm, implementation provided by the author.

We use QuickNet[3] to train our MLP. The final version of the method used 48 subbands and only retained the top 10 principle components of each subband. This

[3] http://www.icsi.berkeley.edu/Speech/qn.html

resulted in the final architecture of the MLP with 480 input units, one hidden layer with 1000 hidden units, and an output layer with 70 units. The training data was split 70% for the training and 30% cross validation of the MLP.

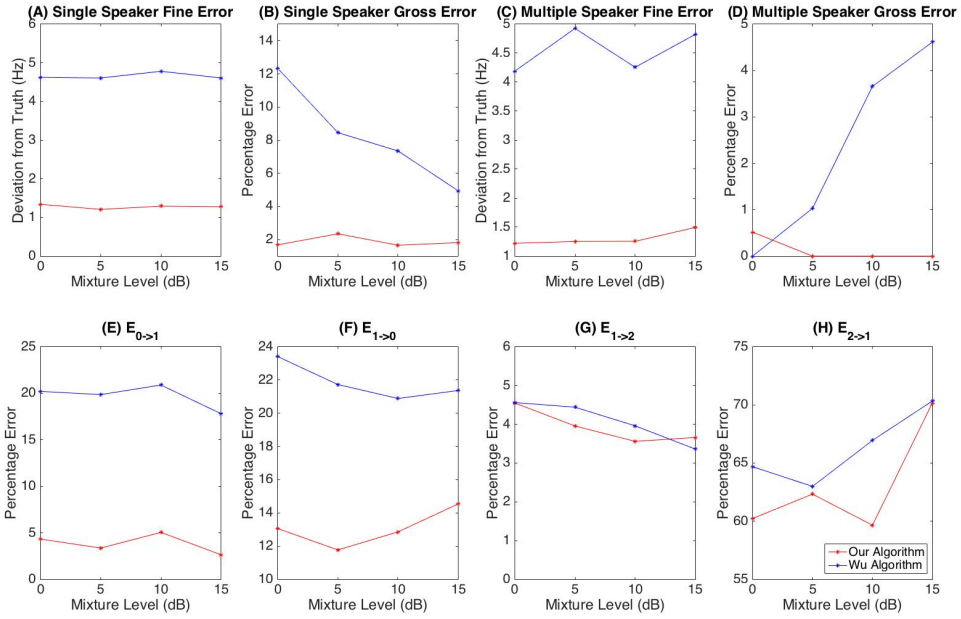

Figure 3: Experimental Results computed over all testing data, Single Speaker Fine Error (plot A), Single Speaker Gross Error (plot B), Multiple Speaker Fine Error (plot C), Multiple Speaker Gross Error (plot D), $E_{0\to1}$*(plot E),* $E_{1\to0}$*(plot F),* $E_{1\to2}$*(plot G),* $E_{2\to1}$*(plot H).Our algorithm is plotted in red and the Wu algorithm in blue.*

Results

Our results indicate an increase in performance over the Wu algorithm in a few of the error metrics we have defined above. We have improved performance in our ability to classify unvoiced and voiced frames (plots E and F) and our pitch state estimate within correctly classified frames deviates less from the ground truth in both single and multiple speaker cases (plot A through D). In addition, the amount of fine and gross error (plots A through D) resulting from our method remain fairly consistent as one speaker becomes more dominant over the other in the mixture, suggesting that our method is robust to the degenerate case we have outlined.

The boost in performance our method received over the Wu algorithm could be in part attributed to the neural net's ability to learn features that aided in the estimation of multiple pitch tracks that the peak selection process could not access. We speculate that by training the neural net to estimate multiple pitch tracks simultaneously, we more accurately estimated the interferer's pitch track and thus reduced the overall error.

Discussion

It is often ambiguous as to where the boundaries of voiced and unvoiced segments are in an utterance and this makes it difficult to correctly label frames on the boundary. The mislabeling of the data contributes to the error in a two-fold manner. First, mislabeling causes error in the training of our classifier, making the classifier weaker. Second, mislabeling causes the number of speakers to be incorrectly counted in the grading of MPDAs being tested. Speaker error is very susceptible to mislabeled ground-truth pitch tracks. This effect can be observed in the top pane of Figure 2 on the starts and ends of voiced track segments.

There are ambiguous situations that represent a large source of multiple speaker error. The two main situations include overlapping pitch tracks from different sources and the situation in which pitch tracks from different sources interleave or pass through each other. It has been observed that when the pitch tracks overlap the classifier will misclassify two speakers as one speaker. In the case where the speaker's pitch tracks interleave, the decoding is ambiguous as to which track belongs to which speaker. These sources influence both speaker error and pitch track error.

Concluding Remarks

This research experiments with the ability of neural networks to perform multi pitch tracking. They provide a succinct framework for the determination of pitch tracks. Future work would include more complex Markov models to handle more than two speakers and to resolve the pass through ambiguity. Different network architectures could also be used to integrate the presence of temporal information into estimating the pitch posterior state distributions to aid the formation of continuous pitch tracks.

Acknowledgments

We would like to acknowledge Professor Dan Ellis (Columbia University Department of Electrical Engineering) for his research guidance in the areas of speech separation and pitch determination algorithms and for a publically available open-source version of the SAcC method.

References

Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press.

Wu, M., Wang, D., & Brown, G. J. (2003). A multipitch tracking algorithm for noisy speech. *Speech and Audio Processing, IEEE Transactions on*, *11*(3), 229-241.

Lee, B. S., & Ellis, D. P. (2012). Noise Robust Pitch Tracking by Subband Autocorrelation Classification. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Rabiner, L., Cheng, M. J., Rosenberg, A. E., & McGonegal, C. A. (1976). A comparative performance study of several pitch detection algorithms. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *24*(5), 399-418.

Plante, F., Meyer, G. F., & Ainsworth, W. A. (1995). A pitch extraction reference database. *Children*, *8*(12), 30-50.

Bagshaw, P. C., Hiller, S. M., & Jack, M. A. (1993). Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, *93*.