Pakistan Journal of Science (Vol. 67 No.4 December, 2015)

PITCH ESTIMATION FRAMEWORK FOR SPEECH SEGREGATION USING COCHLEAGRAM MORPHING

M. J. Khan and H. A. Habib

Department of Computer Engineering, University of Engineering and Technology, Taxila, Pakistan.
Corresponding Author E-Mail: adnan.habib@uettaxila.edu.pk

ABSTRACT: Computational auditory scene analysis (CASA) has significant role in speech segregation from monaural audio mixtures and generally a measure for performance of speech recognition systems. Pitch estimation has a substantial role in performance of CASA systems. This study presents a novel pitch estimation framework for speech segregation from monaural audio mixtures using cochleagram morphing. The proposed framework takes the rough estimation of target pitch from given audio mixtures containing speech and background interferences. Discrete set consisting morphed versions of cochleagram is obtained using k-Means clustering. The estimated pitch values are improved by validating and smoothing them to morphed cochleagram. Measure of refined estimated pitch contours along with harmonicity and temporal continuity are used to segregate target speech. The proposed framework produced 83.13% accuracy for MIR-1k dataset which is considerably higher than the existing methods.

Keywords: Pitch range estimation; Source separation; Computer auditory scene analysis (CASA); k-means; Spectral Peaks.

(Received 14-08-2015 accepted 15-11-2015).

INTRODUCTION

Acoustic signals reaching our ears contain sound waves from multiple sources and reflections from a multitude of objects in the environment. Human brain is sophisticated enough for separately perceiving different sounds from different sources that have been combined together. The multi-source sound detection and recognition is trivial for human hearing. On the contrary, the same task for computing system is extremely daunting. It is still a major challenge to separate sounds from different sources and interfering noises. The cocktail party problem is a classic example of multi-speech processing (Haykin and Chen, 2005). A study has been reported by (Cherry, 1953) about the objective study on cocktail party problem with monaural and binaural listeners. As a result, the theory of computational auditory scene analysis (CASA) evolved with the aim to develop machine systems to achieve sound source separation by exploiting fundamental perceptual principles (Bregman, 1994). Tremendous growth in CASA and sound separation study has been carried in recent years. It has been primarily fueled by the widespread realization that automatic speech and speaker recognition methods lack the ability to handle multi-stream complex auditory signals (Mehla and Aggarwal, 2014). Human voice is principally made up of pitched sounds having different frequencies of concurrent overtones originating from the fundamental frequency F0 (Carroll *et al.*, 2011).

Speech segregation algorithms utilize the harmonic structure of the singing voice (Micheyl and Oxenham, 2010). At first stage, the singing pitch from the song mixtures are extracted as they play role in subsequent separation. Singing voice in a song is a smooth time varying function as compared to musical instruments. A study reported voice separation from music accompaniment based on pre-dominant pitch from vocal segments and computational auditory scene analysis (CASA) perceptual cues (Li and Wang, 2007). The method has been further extended by (Hsu and Jang, 2010) to separate unvoiced spectral components using spectral subtraction. A study reported tandem algorithm estimating the pitch of singing voice and separating the same by improving the pitch estimates iteratively (Hu and Wang, 2010). Such algorithms are referred to as pitch based inference methods. However, the increase in number of auditory sources creates problems for pitch based inference methods.

Statistical techniques such as Principle Component Analysis (PCA), Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF) etc. have been also used for speech separation from multi-stream audio mixtures (Cichocki *et. al.*, 2009; Ozerov and Févotte, 2010; Smaragdis, 2007; Zhu *et. al.*, 2013). These methods blindly try to separate auditory scenes to respective sounds producing sources. Spectrogram factorization-based method applies NMF to decompose the mixture spectrogram into a set of components (Cichocki *et al*., 2009). These group components are used to represent different sound sources. Experiments conducted by (Smaragdis, 2007) reveal that vocals from different musical instruments can effectively be differentiated by limited number of accompanying instrumental sounds. However, spectrogram factorization is immensely challenging when the number of accompanying instruments increases and thereby performance decreases drastically.

Many algorithms have been proposed that fuse pitch-based outcomes with statistical techniques of audio source separation. It has been pointed out that first extract the harmonic structure of singing voice based on fundamental frequency and then feed it to NMF on the residual spectrogram to learn an accompaniment model to separate vocals (Virtanen *et al.*, 2008). The harmonic structure of voiced speech is determined by NMF and one factor is obtained for each discrete pitch value (Sha and Saul, 2004). The resulting dictionary is then used for multi-pitch estimation using nonnegative de-convolution and the contribution of each harmonic template to the observed short-time spectrum is evaluated. Harmonic sources as well as their respective pitch values are retrieved from contribution templates greater than pre-defined threshold.

In the same context many researchers have proposed merging of the outcomes from trained structure of musical instruments with statistical techniques (Raj *et al.*, 2007). Such spectrogram factorization-based singing voice separation techniques have considerably improved the source separation process. These algorithms first train a set of accompaniment spectra from manually-labeled non-vocal segments based on probabilistic latent component analysis (PLCA). Spectra of the singing voice is then computed from the song mixture by keeping fixed to the accompaniment spectra. Another separation method applies the pre-detected non-vocal segments to adapt an accompaniment model for each song mixture (Ozerov and Févotte, 2010). The common point between the above two methods is that both need a significant amount of non-vocal segments in their training data. However, these schemes are limited to the songs of specific structure. The Techniques using background model to extract voice spectra usually require a large repository of template models. Slight variations in background cause significant drop in the performance.

Resolution of time-frequency representation targets to determine the nature of instrumental sounds. Short time windows that exhibit low frequency resolution are good representation of percussive instrumental sound. On the other hand, the longer time windows represent harmonically tuned instruments more precisely**.** Based on time-frequency resolution, Tachibana *et al.*(2014) devised a two-stage method for singing voice enhancement. They used harmonic/percussive sound separation technique (HPSS) on mixture spectrograms with high and low frequency resolutions respectively, the harmonic and percussive elements of the accompaniment can be filtered out from the song mixture, leaving out the vocals. (Fitz and Gainza, 2010) used the same framework but replaced HPSS with a median filtering-based harmonic and percussive separation method. Compared with other vocal separation methods, the multi-stage methods reported by (Tachibana *et al.*, 2014) are more flexible. They require neither pitch detection nor prior training, and do not make assumptions on the repetition and low-ranking of background music.

In this study the focus was on segregation of speech acoustics by improving pitch estimation results with CASA cues from a monaural audio mixture. Rough pitch estimation results have been iteratively improved by incorporating morphed versions of mapped cochleagram. The improved pitch values and CASA cues segregate the target speech from audio mixtures containing speech with background interfering noises. The framework needs no prior information or training about specific sources and hence works efficiently in diverse situations.

METHODS AND MATERIALS

This study proposed pitch estimation frame for speech segregation by confining fundament frequency (F0) using morphed versions of cochleagram. The building blocks of this proposed framework are shown in figure 1.

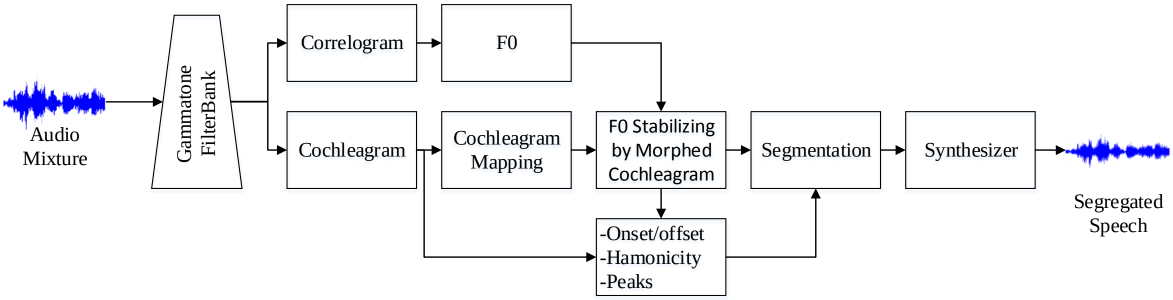


Figure 1. Proposed System Diagram

Audio mixture consisted voices from multiple sources. This audio mixture was fed into gammatone filterbank. A gammatone filterbank is mapping of psycho-physical and physio-logical observations of human auditory periphery (Patterson *et al.*, 1987). The gammatone filter is an impulse response of the product of gamma distribution and sinusoidal tones as stated in Eq. (1)

$$g(t) = at^{n-1}\exp(-2\pi b\, ERB(f_c)t)\cos(2\pi f_c t + \varphi) \quad (1)$$

The channel wise response of gammatone filters was then divided into 20-ms frames with 50% overlapping between the consecutive frames. It became a time-frequency representation of the input audio mixture and called auditory spectrogram or cochleagram (Wang and Brown, 2006).

Cochleagram was first stage in pitch estimation framework. Spectral peaks were identified in each time slot of computed cochleagram. These spectral peaks presented local maxima in each frame of cochleagram. Forward and backward differences were calculated and fused through logical 'AND' operation for each cochleagram frame to compute local maxima as presented by Eq. (2)

$$cg_{peaks}(t,f) = \begin{cases} 1 & cg(t,f-1) > cg(t,f) < cg(t,f+1) \\ 0 & otherwise \end{cases} \quad (2)$$

Where $cg(t,f-1)$, $cg(t,f)$ *and* $cg(t,f+1)$ *were cochleagram units at time t and frequency f* while $cg_{peaks}(t,f)$ were detected peaks of cochleagram at time *t* and frequency *f*. An objective energy threshold $cg_{threshold}$ was applied to filter out large number of peaks with ignorable magnitudes. This threshold value was determined by local maxima of given frame.

$$cg_{peaks}(t,f) = \begin{cases} 1 & cg_{peaks}(t,f) > cg_{threshold}(t,f) \\ 0 & otherwise \end{cases} \quad (3)$$

Where $cg_{peaks}(t,f)$ represents detected peaks that were above threshold energy value $cg_{threshold}$.

The fundamental frequency, F0, is significant component in case of voiced signal. F0 can be estimated via different techniques. This study applied summary correlogram of Gammatone filter bank response as stated in Eq. (4)

$$acf(n,c,\tau) = \sum_{k=1}^{K-1} a(n-k,c)a(n-k-\tau,c)h(k) \quad (4)$$

In this equation, a*(n,c)* characterized the simulated auditory nerve response for frequency channel *c*, at discrete time *n*, and τ is the time lag. The autocorrelation $acf(n,c,\tau)$ was computed over a window of length K samples, which was shaped by a Hanning window function *h*. Fundamental frequency F0 was estimated by summing the information in the correlogram across frequency channel c. The subsequent summary correlogram is given by

$$sacf(n,\tau) = \sum_{c} acf(n,c,\tau) \quad (5)$$

Where $sacf(n,\tau)$ was unidirectional auto-correlation summary of gammatone filterbank response at discrete time *n* and time lag τ. F0 plays vital role in speech music separation as study suggested by using this single feature (Giannakopoulos and Pikrakis, 2014). There are various techniques to have single or multiple F0s from an audio mixture. This study used the correlogram approach for F0 in which summation of channel wise auto-correlated results are used.

Based on observed corresponding F0, initial and terminal points of vocal tones in audio mixture were predicted. False postulants were avoided F0 pruning. 2D representation of Time-Frequency units in cochleagram were reduced to discrete levels according to desired morphed cochleagram shapes that had to be used in next stage. By subjective analysis, eight discrete levels were defined for mapping. K-Means clustering produced mapped cochleagram. K-means required no prior information to classify data as being unsupervised classification technique. The data was classified into k classes based on nearest intensity values as stated in the Eq. (6).

$$\arg\min_{L} \sum_{i=1}^{8} \sum_{L_i} \|cg - \mu_i\| \quad (6)$$

Where cg was time-frequency (T-F) value in frame, L_i was set of T-F units within specific range depending on mean of T-F units'μ_i. By applying k-means, cochleagram was divided the to eight discrete clusters. After clustering, set of frequencies belonged to low energy values were assigned zero value to avoid small fluctuations. The weak values were filtered and clusters of interest were reduced to seven. Voiced speech harmonic structure required that F0 must be located in highest energy cluster (Hu and Wang, 2010). Cochleagram of seven segments had been analyzed one by one from outer most shell to inner most regions.

The pitch contours were heuristically expanded according to more intensive spectral intensities along with temporal continuity. For more intense T-F units in Cochleagram were morphed into single inner most segment that held highest energy values. These results are propagated to first iteration and then to second iteration according to two segments of mapped cochleagram. In the same way, we included one by one a class with lower spectral energy values.

Spectral peaks provided strongest intensity location in cochleagram by two stage pruning of spectral peaks. First segregated the peaks according to stabilized F0 and its marginally equivalent integer multiples using Eq. (7) and secondly with common fate of spectral representation.

$$cg_{peaks}(t,f) = \begin{cases} 1 & cg_{peaks}(t,f) \approx \text{F0 multiple} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Stabilized F0 by morphed cochleagram and in Eq. (7) F0 was further pruned by filtered spectral peaks. In this step systems traversed the estimated values of pitch if abrupt change occurred then seen nearby spectral peak value and smoothed the abrupt change according to nearby spectral peaks. Finally, harmonic filtering had been applied.

Onsets and offsets correspond to sudden intensity changes and call common fate. Common fate is second most prominent feature after pitch (Bregman, 1994). In this framework, we employed three steps procedure (Hu and Wang, 2007) for getting onset and offset information. In first step audio signal was smoothed; second step identified peaks and valleys and third step applied threshold to filter ignorable peaks and valleys.

The result of stabilized pitch values and their corresponding spectral peaks were used to label T-F unit. A multi scale onset/offset based segmentation algorithm (Hu and Wang, 2007) had been applied that produced segments enclosed by detected onset and offset information. These segments were further filtered according to morphed version of cochleagram for further segmentation.

RESULTS AND DISCUSSIONS

A) Gammatone Filterbank: Gammatone filterbank received input of audio mixture. This audio mixture contained speech and Karaoke version of music. It was sampled at 16 KHz sampling frequency and fed into 256-channel gammatone filterbank with sample frequencies equally spaced from 50 Hz to 5 KHz on the equivalent bandwidth rate scale.

B) Cochleagram: The channel wise response of Gammatone filters was then divided into 20-ms frames with 50% overlapping between the consecutive frames. This scheme provided a time-frequency representation of the input audio mixture, namely auditory spectrogram or Cochleagram (Wang and Brown, 2006). Cochleagram representation of audio mixture was shown in Figure 2.

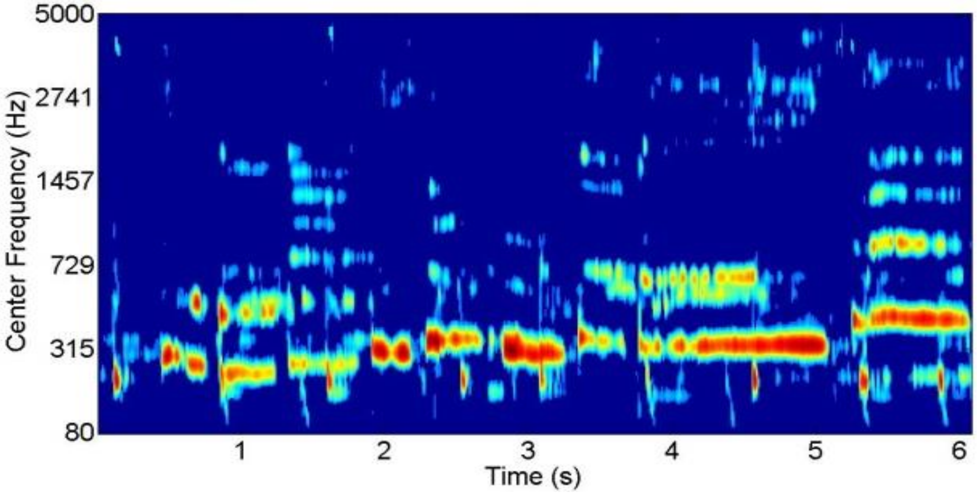


Figure 2. Cochleagram of audio mixture randomly picked from MIR-1k dataset

C) Spectral Peaks: Cochleagram was basically first stage of pitch estimation framework. After the construction of cochleagram, spectral peaks have been found in each time slot. Spectral peaks provided local maxima in each frame of cochleagram. To complete local maxima forward and backward differences were calculated and fused through logical 'AND' operation for each cochleagram frame as shown in the figure 3.

D) Fundament Frequency Estimation: The fundamental frequency F0, was calculated by summary correlogram of Gammatone filter bank response, as shown figure 4.

Summary correlogram gammatone filter bank characterized the simulated auditory nerve response for frequency channels at discrete time. In this summary, peak that occurred at the shortest time lag τ (i.e. 10ms) correspond to fundamental frequency. As per research (Yang, 1996) F0 of human beings lied in the range between 80 Hz to 500 Hz in general. It was further reported that male speakers had in the range 70 Hz to 200Hz, female speakers had the range 140 Hz to 400 Hz and children had the range from 180Hz to 500Hz. It was ensured that resultant F0 must be in the range 80Hz to 500Hz.

F0 plays vital role in speech music separation as this study achieved 88% accuracy using this single feature (Giannakopoulos and Pikrakis, 2014). There are several other techniques based on single or multiple F0s from an audio mixture. It was observed that

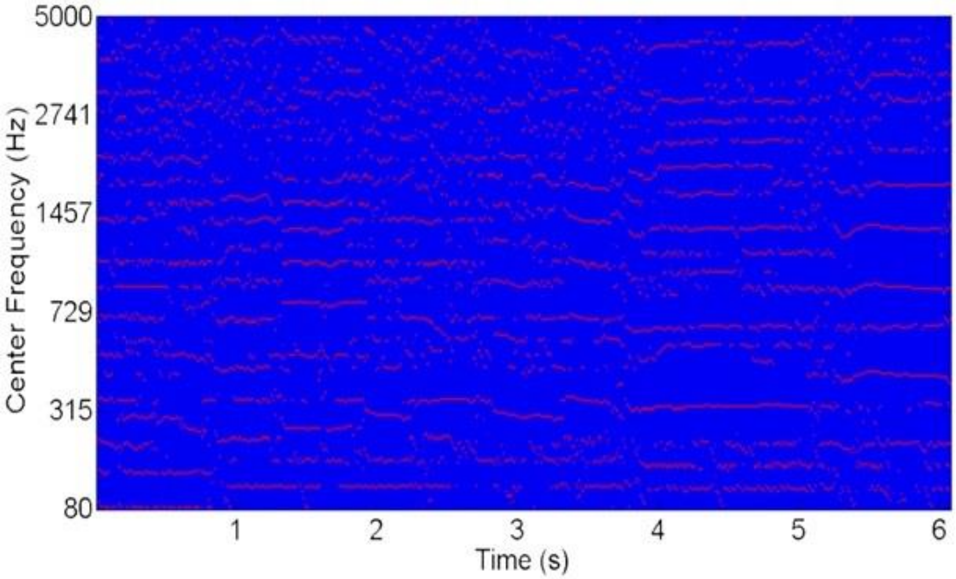


Figure 3. Detected spectral peaks from cochleagram shown in figure 2

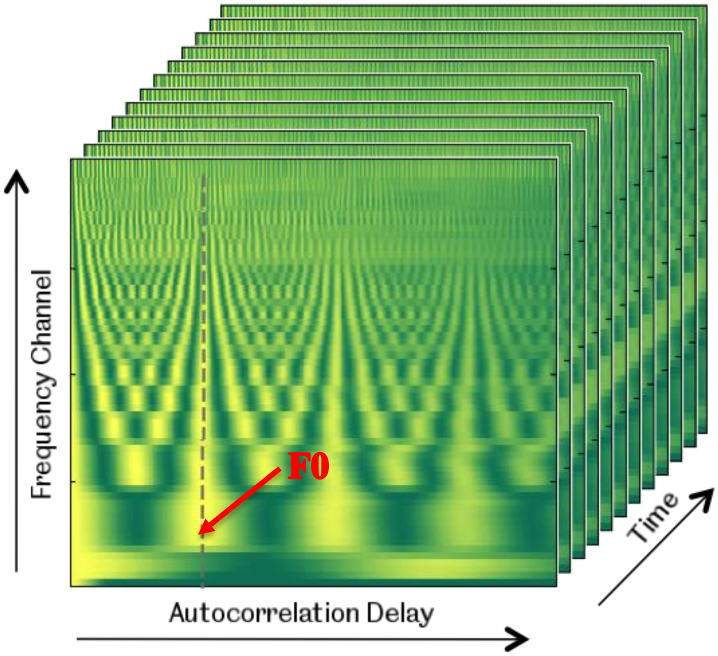


Figure 4. Time lag calculation for F0 from correlogram frames

- F0 responses to harmonical instruments remained constant over course of time.
- F0 against percussive instruments produced no effective result at all.
- F0 changed over the time to vocal sounds.
- F0 exhibited the same as above in response to presence of vocal and harmonical instrument sounds.

Based on observation corresponding F0, it was roughly predicted initial and terminal points of vocal tones in audio mixture while present alone or combined with some pitched instruments. We can afford above results if F0 pruning was considered to avoid false postulants.

E) Cochleagram Mapping: Cochleagram mapping mapped the cochleagram time-frequency units to lesser number of discrete levels to achieve desired morphed Cochleagram shapes used in next stage. By subjective analysis, eight discrete levels for mapping were defined. Then mapped Cochleagram was produced by applying K-

Means clustering. The result of K-Means clustering is shown in figure 5.

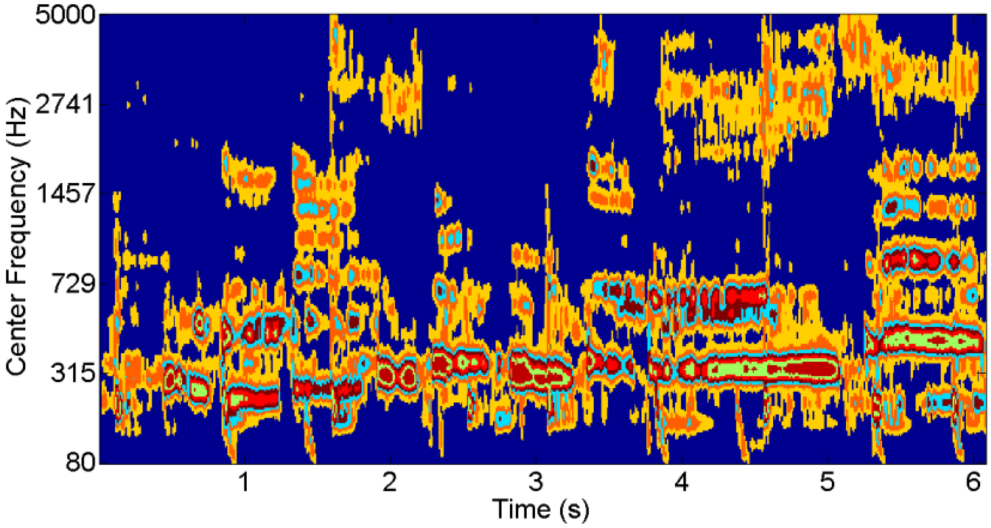
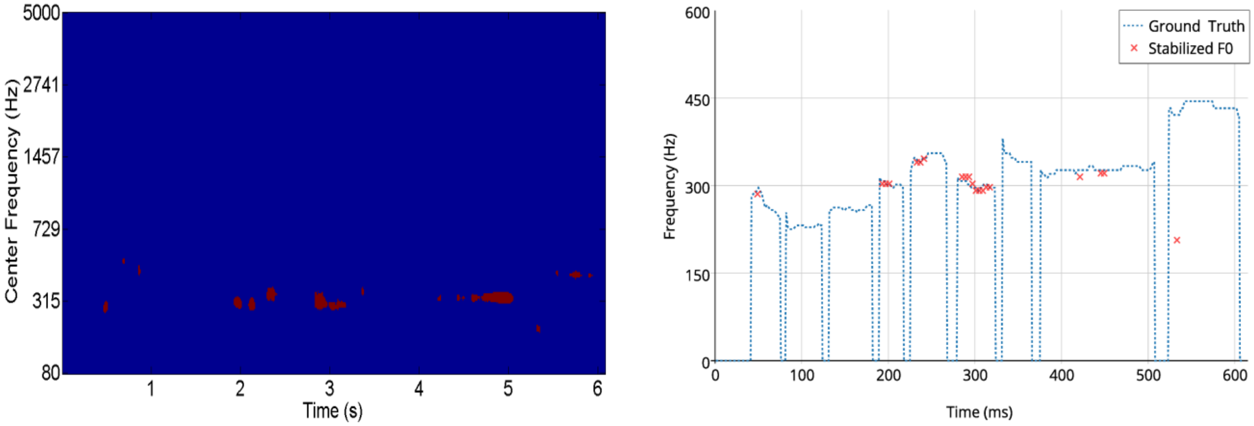


Figure 5. Mapped cochleagram of audio mixture shown in figure 2.

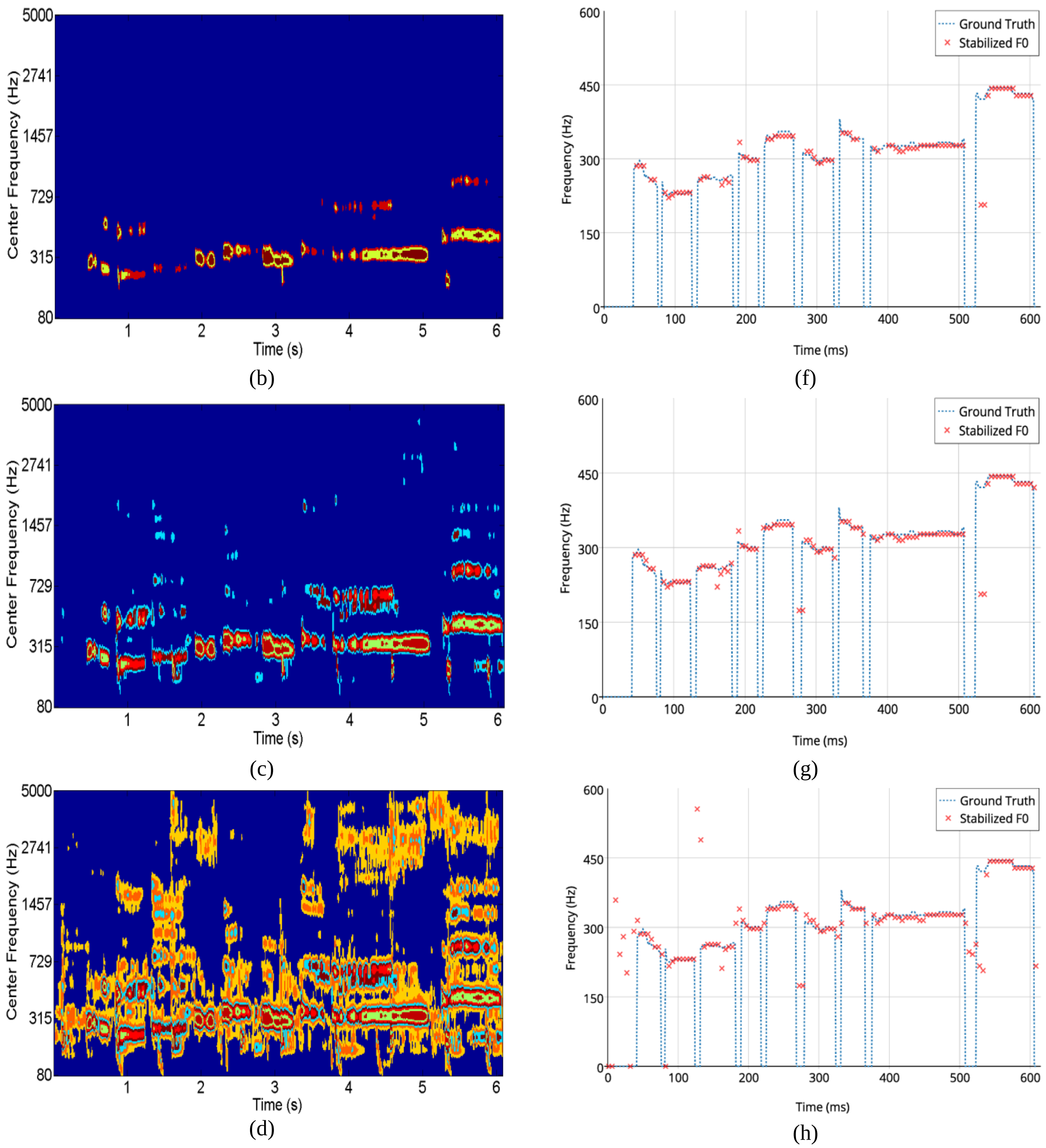
After clustering, set of frequencies belonging to lowest energy values were truncated to zero to avoid small fluctuation and clusters of interest were reduced to seven. This concept was based on harmonic structure of voiced speech with the assumption that F0 must be located in highest energy cluster (Hu and Wang, 2010).

F) Iterative Stabilization of Cochleagram Morphing: Iterative stabilization of cochleagram morphing was achieved by heuristically expanding the pitch contour according to intensive spectral intensities along with temporal continuity. For more intense T-F units, cochleagram was morphed into single inner most segment that held highest energy values. So the morphed Cochleagram overlaid with estimated F0 resulted from auto-correlation of gammatone filter response. Regions that laid with estimated F0 values were declared valid. Based on valid F0 and their corresponding Cochleagram region, this validity grown to whole region by allotting vertical center pixels of that region. Results were propagated from first iteration to second iteration that involved morphing of Cochleagram according to two segments of mapped Cochleagram that were inner most regions belonged to class of highest spectral energy values and its adjacent outer segments that belonged to moderate spectral intensity class with respect to first inner segment. In the same way, one by one a class with lower spectral energy values is included. Figure 6 is an illustration of morphed version of cochleagram and their corresponding F0 stabilization. Inclusion of seven segments of mapped Cochleagram one by one as shown in figure 6(a) higher intensity T-F units L_1 to figure 6(d) lower intensity T-F units L_7 and corresponding to morphed cochleagram, extended F0 validity as shown in figure 6(e) to figure 6(h) to whole cochleagram. Once valid F0 calculated it will be valid for whole vertical axis of Cochleagram until next step.



(a)

(e)



(b) (f)

(c) (g)

(d) (h)

Figure 6. F0 stabilization by incremental cochleagram morphing (a) Morphed with class highest intensity values (b) Morphed with combination of three classes of higher order (c) Morphed with combination of five classes of higher order (d) Morphed with combination of seven classes and (e)-(h) their corresponding stabilized F0.

G) Harmonic Filtering of Spectral Peaks: Spectral peaks provided strongest intensity location in Cochleagram. Two stage pruning have been applied for locating spectral peaks. First peaks were segregated according to stabilized F0. Stabilized F0 by morphed cochleagram was further pruned by filtered spectral peaks. In this step systems traversed the estimated values of pitch if abrupt change occurred then seen nearby spectral peak value and smoothed the abrupt change according to nearby spectral peaks. Harmonic filtering of spectral peaks produced more accurate results as shown in figure 7.

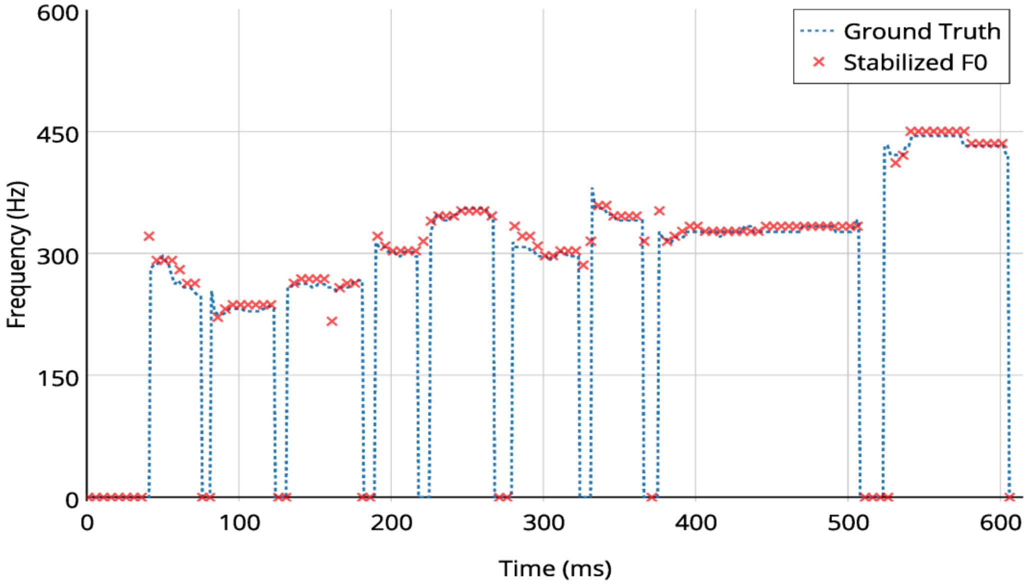


Figure 7. Result of Stabilized F0 after its alignment to filtered spectral peaks

H) Onsets/Offsets Detection: Onsets and offsets correspond to sudden intensity changes and call common fate. Common fate is second most prominent feature after pitch (Bregman, 1994). In this framework, we employed three steps procedure (Hu and Wang, 2007) ofgetting onsets and offsets. In first step audio signal was smoothed; secondly peaks and valleys were identified and then threshold was applied to filter ignorable peaks and valleys.

I) Evaluation of Pitch values: Proposed framework is based on cochleagram morphing based pitch estimation and speech segregation algorithm produced pitch contour for target voiced source and associated binary mask. In this section we evaluated proposed framework for pitch estimation and speech segregation. This research work experimented with MIR-1K dataset publically available (Hu and Wang, 2010). This dataset contained 1000 karaoke version Chinese songs. These songs are mixed with immature singers' voices. The duration each clip ranged from 4 to 13 seconds, and the total length of data set was 133 minutes featuring 19 singers (including 8 females and 11 males). The work summed Karaoke music and singing voice channels of each segment at different SNR levels and treated them as a single channel. The ground truths of the target pitch of clean singing voices were also the part of dataset.

Table: 1. Percentage of correct detection of pitch values at different SNRs, compare against well-known existing methods for pitch extraction.

Input SNR	Proposed	Hu and Wang (2010)	Li and Wang (2007)	Boersma and Weenink (2013)	Wu *et al*. (2003)
-5 dB	67%	69%	44%	33%	51%
0 dB	79%	76%	58%	48%	62%
5 dB	87%	78%	70%	69%	69%
10 dB	90%	82%	80%	82%	70%
15 dB	91%	83%	88%	83%	71%

This research work considered pitch values that were located within morphed structures of cochleagram and had some harmonic nature. The estimated pitch was considered as correct if the difference between estimated pitch and ground truth pitch was less than 5% in Hz. All experiments were performed without any prior knowledge of instruments used in music. The correct rate for singing voiced pitch detection was calculated at different SNR levels. The performance of our proposed framework had been compared with well-known existing approaches as it seen in table 1. The proposed framework achieved 67% accuracy at -5 dB SNR which was better than the methods of Li and Wang (2007), Boersma and Weenink (2013), and 0 dB and higher dBs. The framework has been tested to different levels of dBs to compare the performance with the methods. Significant results have been achieved by the proposed framework. The reason for better results is due to confinement of high intensity units in cochleagram. The cochleagram morphing algorithm for pitch estimation performed very well. Nevertheless, the cochleagram morphing based pitch estimation achieved 83.13% accuracy on average and was robust than competitive performance at different SNRs and it outperformed on higher SNRs.

J) Evaluation of Segregated Speech: In addition to pitch estimation, we evaluated voiced speech segregation performance on a corpus of 100 mixes of voiced speech and intrusions, conventionally used for CASA investigation (Hu and Wang, 2010). These interferences were N0-1 kHz pure tone, N1- white noise, N2- noise burst, N3- cocktail party noise, N4- rock music, N5- siren, N6- telephone, N7- female speech, N8- male speech and N9- female speech. The inferences were considerably diverse structures e.g. N1 white noise corrupt the whole range of frequencies or N3 harmonic nature noise of cocktail part noise. We measured the target speech separation performance by comparing the waveforms of voiced speech signal and the segregated voiced target using Eq. (7) (Hu and Wang, 2010). We used Eq. (7) for the comparison of waveforms in dBs.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n} s^{2}(n)}{\sum_{n}[s(n)-\check{x}(n)]^{2}}\right) \qquad (7)$$

where $\check{x}$ was the resynthesized signal and s(n) was the target voiced speech signal before being mixed with the intrusion. Our system outperformed against intrusions N0-N4 and N6 as compared to (Hu and Wang, 2004) and spectral subtraction (Cooke, 2005) which was the standard method of speech enhancement. The proposed framework performed consistently better when compared against (Hu and Wang, 2004) and spectral subtraction (Cooke, 2005). Comparative SNR based results are shown in figure 8. In average, the proposed system obtained a 17.04 dB SNR gain, which was about 2.11 dB better than Hu and Wang's system and 8.61 dB better than the Spectral Subtraction.

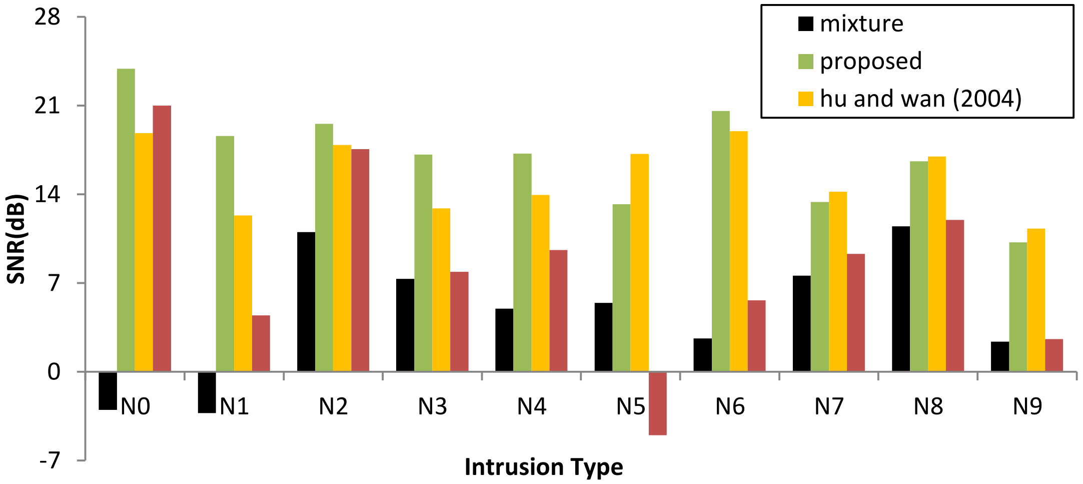


Figure 8. Comparative SNR results for segregated speech and original mixtures for a corpus of voiced speech and various intrusions

Conclusion:

In this paper, a novel framework for voiced speech separation from monaural audio mixture was proposed. The framework applied k-means clustering technique on Cochleagram produced from gammatone filterbank response to map cochleagram to 8 discrete clusters of time frequency units. Mapped Cochleagram was used to make discrete set of morphed Cochleagram versions and stabilized rough pitch estimation by morphed Cochleagram versions from highest intensity values to lowest intensity values. Estimated pitch accompanied with harmonicity and temporal continuity was used to generate labeled binary mask for speech segregation. This study had been evaluated on MIR-1k dataset that contained 1000 karaoke version Chinese songs. The experimentations showed that estimated pitch contours were closer to true pitch. The framework outperformed other state of the art methods. The proposed framework estimated significantly better pitch contours for both structured and unstructured background noises.

Acknowledgments: This work had been supported by the Directorate of Advanced Study and Research (ASR&TD) University of Engineering & Technology Taxila.

REFERENCES

Bregman, A. S. (1994). Auditory scene analysis: The perceptual organization of sound: MIT press.

Boersma, P. and D. Weenink (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/

Carroll, J., T. Stephanie, and Z. Fan-Gang (2011). Fundamental frequency is critical to speech perception in noise in combined acoustic and electric hearinga. The Journal of the acoustical society of America, 130(4), 2054-2062.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America, 25(5), 975-979.

Cichocki, A., R. Zdunek, A. H. Phan, and S. Amari (2009). Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation: John Wiley & Sons.

Cooke, M. (2005). Modelling auditory processing and organisation (Vol. 7): Cambridge University Press.

Fitz, G. D. And M. Gainza (2010). Single channel vocal separation using median filtering and factorisation techniques. ISAST Transactions on Electronic and Signal Processing, No. 1, Vol. 4, 2010, pages: 62 – 73.

Giannakopoulos, T., and A. Pikrakis (2014). Introduction to Audio Analysis: A MATLAB® Approach: Academic Press.

Haykin, S., and Z. Chen (2005). The cocktail party problem. Neural computation, 17(9), 1875-1902.

Hsu, C. L., and J. S. R. Jang (2010). On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. Audio, Speech, and Language Processing, IEEE Transactions on, 18(2), 310-319.

Hu, G., and D. Wang (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. Neural Networks, IEEE Transactions on, 15(5), 1135-1150.

Hu, G., and D. Wang (2007). Auditory segmentation based on onset and offset analysis. Audio, Speech, and Language Processing, IEEE Transactions on, 15(2), 396-405.

Hu, G., and D. Wang (2010). A tandem algorithm for pitch estimation and voiced speech segregation. Audio, Speech, and Language Processing, IEEE Transactions on, 18(8), 2067-2079.

Li, Y., and D. Wang (2007). Separation of singing voice from music accompaniment for monaural recordings. Audio, Speech, and Language Processing, IEEE Transactions on, 15(4), 1475-1487.

Mehla, R., and R. Aggarwal (2014). Automatic Speech Recognition: A Survey. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), 3(1), pp: 45-53.

Micheyl, C., and A. J. Oxenham (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. Hearing research, 266(1), 36-51.

Ozerov, A., and C. Févotte (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. Audio, Speech, and Language Processing, IEEE Transactions on, 18(3), 550-563.

Patterson, R., I. Nimmo-Smith, J. Holdsworth, and P. Rice (1987). An efficient auditory filterbank based on the gammatone function. Paper presented at the a meeting of the IOC Speech Group on Auditory Modelling at RSRE.

Raj, B., P. Smaragdis, M. Shashanka, and R. Singh (2007). Separating a foreground singer from background music. Paper presented at the International Symposium on Frontiers of Research on Speech and Music, Mysore, India.

Sha, F., and L. K. Saul (2004). Real-time pitch determination of one or more voices by nonnegative matrix factorization. Departmental Papers (CIS), 168.

Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. Audio, Speech, and Language Processing, IEEE Transactions on, 15(1), 1-12.

Tachibana, H., N. Ono, and S. Sagayama (2014). Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 22(1), 228-237.

Virtanen, T., A. Mesaros, and M. Ryynänen (2008). Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. Paper presented at the SAPA@ INTERSPEECH.

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

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

Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. Journal of Phonetics, 24(2), 245-261.

Zhu, B., W. Li, R. Li, and X. Xue (2013). Multi-stage non-negative matrix factorization for monaural singing voice separation. Audio, Speech, and Language Processing, IEEE Transactions on, 21(10), 2096-2107.