International Journal of Scientific & Engineering Research, Volume 5, Issue 9, September-2014
ISSN 2229-5518

Analysis and synthesis of sinusoidal noise in monaural speech using CASA

FathimaC.M
M.tech.Applied Electronics
Ilahia college of engineering and technology
Kochi,India
iqbal.fathima.fathima@gmail.com

Khadeeja mol .K.U
Asst. Professor Electronics & communication engineering
Ilahia college of engineeting &tech
Kochi, India
khadeejamol@icet.ac.in

Abstract:**CASA is the technique used to segregate a target speech from a monaural mixture. This article propose a technique to separate the sinusoidal noise from monaural mixtures. Many sounds are there that are important to humans are having pseudo-periodic structure over a particular period /stretch of time. Where this fixed period is typically range of 100Hz-5KHz which gives the corresponding pitch percept. The systematic evaluation of this algorithm gives a tremendous and noticeable improvement in noise segregation.**

Keywords:monaural speech segregation, CASA, sinusoidal noise analysis

1.Introduction

Human beings are capable to distinguish and track various noisy environments, while this remains as a big challenge to computers. 'Auditory scene analysis' written by Bergman published in 1990was the first explained the perception and analysis of complex acoustic mixtures which inturns lead to the invention of the computational model, CASA(computational auditory scene analysis).

Human auditory system is simulated and processed by a CASA as similar to the human auditory perception. This has two stages: first is the segmentation and then grouping. The input signal is decomposes to sensory segments in segmentation stage and the signals that are likely came from same source is grouped together as 'target stream'. CASA is capable of dealing with monaural speech segregation efficiently and its getting more efficient in time by time.

Brown and cook who proposed the CASA system which employs maps of many of the auditory features from the cochlear model of speech segregation. And a priori knowledge of input signal is does not require for this system but have some limitations ie, it cannot handle sequential grouping problem effectively and often leaves missing parts in the segregated speech [4].

CASA model for voiced speech segregation is proposed by Wang& Brown[3,5]and is based on oscillatory correlation. For this it uses harmonicity and temporal continuity as major grouping cues. And this implementation is able to recover most of the target speech back , but was unable to get high frequency signals back.

Hu&Wang[6,7] proposes the system for the voiced speech segregation and it is a typical monaural system; and this groups the unresolved and resolved harmonics separately. And in [8] for pitch estimation an improved tandem algorithm is provided.

Multi scale offset and onset analysis is analysis is employed for the unvoiced speech segregation in Hu-Wang system.

Acoustic phonetic features are used in classification stage after voiced speech segregation for distinguishing unvoiced segments from interference [10, 11].

This article proposes an improved and advanced system for sinusoidal noise analysis and synthesis by using computational auditory scene analysis.

For the periodic signals, which can be approximated by the sum of sinusoids and whose frequencies are the integer multiple of the fundamental frequency and the magnitude and phase can be uniquely determined to match the signal called Fourier analysis. Spectrogram is one of the manifestation which shows short time Fourier transform magnitude as a function of time.

A series of normally horizontal , uniformly spaced energy ridges is revealed by a narrowband spectrogram which is correspond to the sinusoidal Fourier component of harmonics which is an equivalent representation of the sound waveform.

To represent each of these ridges explicitly and separately, as a set of frequency and magnitude values is the key idea and is the aim of sine wave modelling.

In monaural speech segregation response energy feature plays an important role in initial segmentation. In formal CASA system T-F (Time-Frequency)'s response energy was taken as a constant value and is used as the threshold which was less efficient since the intrusions are unknown.

The binary mask map is constructed after further grouping and unit labeling. The scattered and broken auditory elements present in the binary mask will produce unwanted fluctuations and utterance which in turns degrade the quality of the resynthesized speech.so in [6] Hu- Wang system includes a smoothing stage in order to avoid this unwanted fluctuation by removing the segments shorter than 30ms and so on.

2. SINE WAVE ANALISIS

Sine wave analysis is a quite simple concept. As shown in the spectrogram ,from the short time Fourier transform, frequency and the magnitude of the spectral peaks at

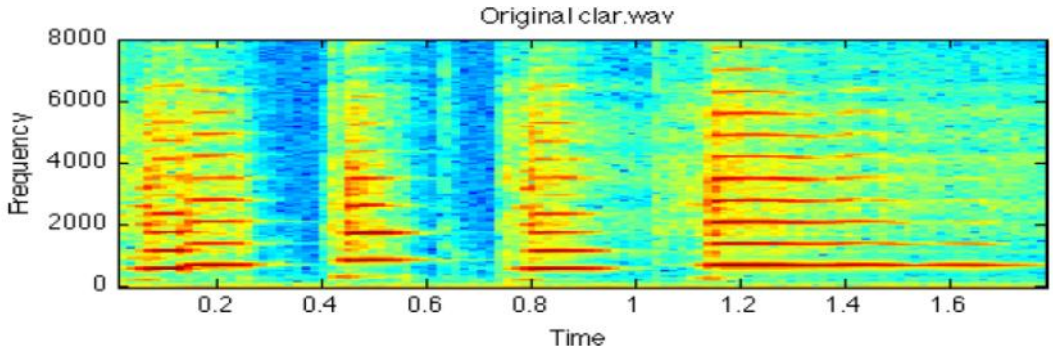


Figure 1. spectrogram representation of sine wave analysis

each time step is find out and thread them together and the representation will be obtained.

It get complicated because of a couple of reasons. First one is difficulty in picking up peaks. Also resolution of STFT is typically not all that good. So the need of interpolating the maximum in both frequency and magnitude arises.

3. SYSTEM DESCRIPTION

Figure 2 represents the system for monaural speech segregation based on CASA. And comparing with other segregation since it is using morphological image processing so an additional smoothing stage is added to improve the initial segmentation stage.

3.1 Basic periphery processing

In the initial stage 128 channel gamma tone filter banks and a simulation of neuro mechanical transduction of inner hair cells is used to model auditory periphery system. The input signal is decomposed into T-F domain by passing through the auditory periphery model. The psychological observation of auditory periphery will provide the gamma tone filters and it's the standard model of cochlear filtering. The impulse response of the gammatone filer is given by,

$$g(t)=\begin{cases} t^{(l-1)}\exp(-2\pi bt)\cos(2\pi ft) & t \geq 0 \\ 0 \quad ; \quad else \end{cases} \quad (1)$$

a low pass filter is used to extract the response energy feature of every channel [6]. The output is represented as *h(c,n).*

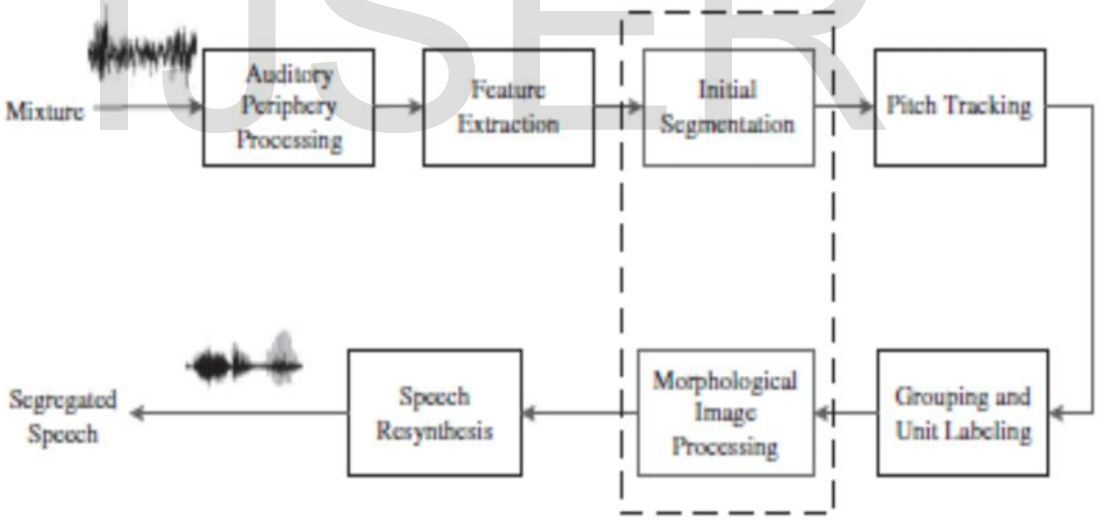


FIGURE 2. Block diagram for improved monaural speech segregation based on CASA

3.2 Feature Extraction

1.Correlogram: Auto correlation of inner hair cell response *h(c,n)* in the T-F domain is used to construct the correlogram.

$$A_H(c.m,\tau) = \frac{1}{N_c}\sum h\,(c,mT-n)h(c,mT-n-\tau) \quad (2)$$

Where, c is the order of the channel , m is the time frame, N_c is the no.of samples in a frame of 20 ms.

2. Cross channel correlation: it indicates wheather the filter responds to the same target. And is calculated as,

$$C_H(c,m) = \sum_{\tau=0}^{L-1} \grave{A}_{H\,(c,m,\tau)}\, \grave{A}_{H\,(c+1,m,\tau)} \quad (3)$$

Where L is the sampling no. and $\grave{A}_H$ is A_H normalized to zero mean and unity variance.

3. Response Energy: The response energy is the correlogram A(c,m,0) when $\boldsymbol{\tau}$=0.

4. Onset/Offset detection: sudden intensity change is expressed in terms of onset and offsets.

3.3 Initial Segmentation

It comprise of two parts one is voiced and another one is unvoiced speech segregation. onset and offset method is used for unvoiced segmentation where as the voiced segmentation is based on extracted features.

Comparing the background noise and targeted speech the later has more stronger response energy of T-F units. The energy features A(c,m,0)and the cross channel correlation feature C(c,m) is used for the estimation of estimated target and is as follows[6]

$$\begin{cases} A(c,m,0) \geq \theta_{Hc} \\ C(c,m) \geq \theta_c \end{cases} \quad (4)$$

Where, θ_c is the constant and is 0.985 [5] and θ_{Hc} is the threshold for effective target energy and,

$$\theta_{Hc} = \frac{1}{M.\alpha}\sum_{m=1}^{M} A(c,m,o) \quad (5)$$

Where M= total no. of frames in a single channel and α is the constant which decides the threshold and is approximated to 1.2.

3.4 Pitch Tracking

For the CASA system the tracking and detection of pitch in complex environment is quite difficult and seems to be challenging .

But the use of tandem algorithm makes things easier ,it can track many pitch contours and can efficiently handle the multi talker problem. For this, primarily the pitch estimation should be complemented.

From the segmented units the units which with strong energy and high cross channel correlation are taken likely from the target speech and these are called as active units. And the estimated target pitch is calculated as ,

$$\tau_{s,1}\,(m) = \arg max \sum_c L_0(c,m).sgn(P(H_o{:}r_{cm}(\tau)) - \theta_p) \quad (6)$$

$$\text{Where sgn(x)} = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < o \end{cases} \quad (7)$$

Similarly the second pitch

$$\tau_{s,2}\,(m) = \arg max \sum_c L_2(c,m).sgn(P(H_o{:}r_{cm}(\tau)) - \theta_p) \quad (8)$$

Where $L_n(c,m)$ is the mask re estimated.

3.5 Grouping and Unit Labelling

In this stage streams are formed by grouping T-F units and these groups are labelled in to target speech and back ground noise. Segregation is needed for both voiced and unvoiced speeches. For the non-speech interference tandem algorithm is used for the voice speech segregation. And if the intrusion is another speech then grouping is performed by analyzing the pitch contour

3.6 Morphological Image Processing

Intrusions can be suppressed by using proper morphological image processing; It is performed by removing the unwanted particles and complementing the broken auditory elements thereby enhancing the segregated speech

The proper dilation and erosion is fundamental in morphological image processing.

(i)Dilation: It is the process that "thickens" or "grows" the object in a binary image. The thickening is controlled by a structuring element B.

Let B is the structuring element and A is the mask, $\acute{B}$ is the reflection set and is defined as ,

$$\acute{B} = \{c | c = -b \; for \; b \in B\} \quad (9)$$

Let $(B)_z$ is the translation of B by the point z=(z_1, z_2) and the dilation

$$(B)_z = \{c | c = b + z \; for \; b \in B\} \quad (10)$$

And the dilation is,

$$A \bigoplus B = \{z | (\acute{B})z \cap A \neq \emptyset\} \quad (11)$$

(ii) Erosion: It is the process ot "Thins" or "Shrinks" the object in a binary image.

$$A \ominus B = \{z | (\acute{B})z \cap A^C \neq \emptyset\} \quad (12)$$

Mask smoothing is carried out by using morphological image processing. In this stage active elements are considered to have similar periodicity pattern. And , the smoothing extend is defined by the simulating element B

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (13)$$

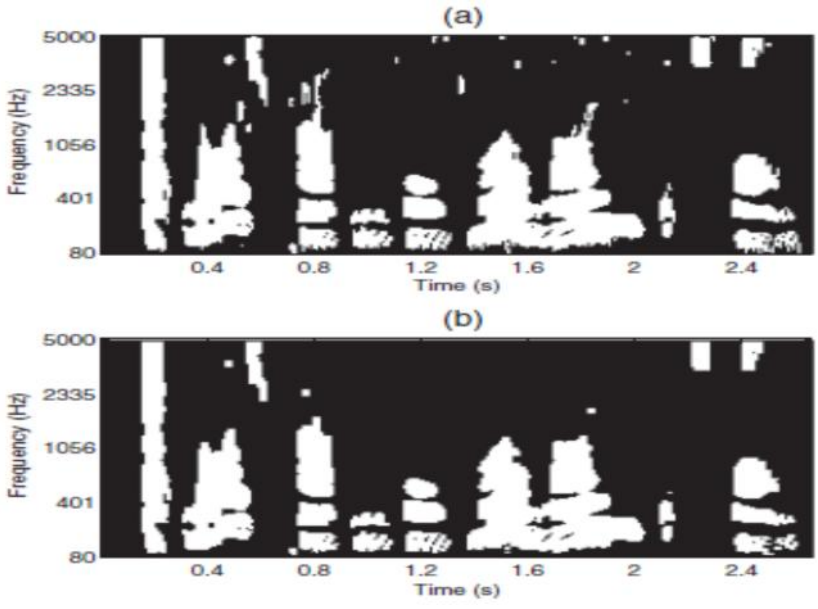


Fig 4. An illustrative example of pruning the mask (a)original mask (b) the mask $\acute{A}$ after pruning

pruning is the process that is used to remove the isolated particles and smooth the spurious salience in the segments in the obtained mask. And is represented as,

$$A' = (A \ominus B) \oplus B \quad (14)$$

Where as complementing is

$$C_{10W} = (A'_{10W} \oplus B) \square B \quad (15)$$

Is applied after pruning on the broken auditory elf in the low frequency range. For high frequency range residual interference energy distributed. For high frequency is complimentary is supplied unnecessary mode will brought in to segregated speech

3.7 Re synthesis

Segregated speech is resynthesized after smoothing stage. While analyzing the sine waves, resynthesize is based on analysis by using simple sine wave oscillator bank.

And tracking and resynthesizing harmonic peaks with sinusoids works pretty well. But some energy was not reproduced such on breath noise.

3.8 Residual Extraction

Resynthesizing and tracking of the harmonic peaks with sinusoids worked pretty well. Even then some energy is not reproduced, like the breath noise because it didn't results in any harmonic peaks. Thus the by subtracting the resynthesized signal from original signal will results in the final signal.In practice it will not work if we are not careful to make the frequencies ,magnitude and phase of the reconstructed sinusoid exactly match the original.

4.Evaluation And Comparison

For validating the effectiveness of a proposed method it is necessary to have a comparison study with existing system. The data base consist of 170 mixtures obtained by mixing 17 intrusions at various SNR levels. The original utterance are selected randomly from the TIMIT data base. The Fs is selected as 16KHz. The intrusions selected are, N1,white noise; N2, rock music ; N3,siren ;N4, telephone; N5, electric fan; N6, alarm clock;N7 traffic noise;N8bird chirp with water flow;N9 wind noise ;N10 rain;N11 cocktail party ;N12 crowd noise at a play ground;N13crowd noise with music ;N14 crowd noise with clap;N15 babble noise;N16 male speech; N17 female speech. In the last two cases the interference is much weaker than the target utterance.

The average SNR is selected by using the equation

$$SNR = 10 log_{10} \frac{So(n)^2}{(S'(n) - So(n))^2} \quad (16)$$

So(n) is the original speech and S'(n) is the segregated speech. system.

The better the performance, lower the P_{EL} and P_{NR} will be and vice versa.

The result of the comparison is given in the following tables includes final SNR result, comparison of $P_{EL,} P_{NR} P_{ESQ}$

Table 1.Final SNR results

Intrusion	Morphological image processing	Proposed system
N1	14.88	15.08
N2	10.08	11.2
N3	13.72	13.9
N4	13.46	14
N5	10.18	12.3
N6	13.37	14.60
N7	7.13	9.2
N8	12.64	13.5
N9	8.94	10.5
N10	13.45	15.6
N11	11.62	11.82
N12	12.01	12.72
N13	10.94	13.5
N14	9.27	10.6
N15	9.09	11.87
N16	15.56	16.7
N17	11.15	11.85
average	11.62	12.87

Table 2. Final P_{EL} P_{NR} result

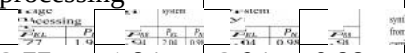
Intrusion	Morphological image processing		Proposed system	
				
N1	2.77	1.94	2.04	0.98
N2	5.03	3.95	4.13	2.73
N3	5.92	4.08	5.02	3.97
N4	2.11	4.73	1.1	3.58
N5	1.93	1.01	1.04	0.97
N6	1.70	1.88	1.23	1.12
N7	5.21	6.23	4.87	5.57
N8	5.77	0.69	4.93	0.72
N9	3.95	4.72	3.92	3.69
N10	2.06	0.42	2	0.45
N11	3.22	1.89	3.11	1.57
N12	2.01	2.19	1.98	2.02
N13	3.14	2.25	2.87	1.56
N14	4.96	3.62	4.21	2.77
N15	3.83	3.72	2.59	3.05
N16	1.84	1.87	1.83	1.57
N17	4.06	6.10	4.02	5.11
average	3.50	3.02	2.99	2.43

Table 3. PESQ Result

intrusion	Morphological image processing	Proposed system
N1	2.272	2.381
N2	1.729	1.736
N3	2.30	2.29
N4	2.33	2.360
N5	1.775	1.815
N6	1.821	1.834
N7	1.540	1.551
N8	1.671	1.693
N9	1442	1.499
N10	1.518	1.618
N11	1.673	1.800
N12	1.668	1.645
N13	1.468	1.490
N14	1.732	1.753
N15	1.492	1.503
N16	1.714	1.705
N17	1.654	1.652

5. CONCLUSION

This article concentrates on the synthesis and removal of sinusoidal noise from monaural speech. The segregation is carried out with the aid of CASA. In this an improved threshold selection results in the better performance. While analyzing the SNR , P_{EL}, P_{NR} it is clear that the proposed system has been improvement in terms of reduction in noise and cutting the energy loss.

Abbreviations:

CASA-computational auditory scene analysis, IBM-ideal binary mask, SNR-signal to noise ratio, PESQ- perceptual evaluation of speech quality.

References:

1.Improved monaural speech segregation based on computational auditory scene analysis.WangU,Chen Niang,Yaun Wenhao

EURASIP journal on audio,speech and Music processing2013
2) .A Bregman, *Auditory Scene Analysis*. (MIT Press, Cambridge, MA, 1990)
3) .G Brown, M Cooke, Computational auditory scene analysis. Comput Speech Lang. **8**, 297–336 (1994)
4) D Wang, G Brown, Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Netw. **10**(3), 684–697 (1999)
5)G Hu, D Wang, Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Netw. **15**(5), 1135–1150 (2004)
6.) G Hu, D Wang, An auditory scene analysis approach to monaural speech segregation, *Topics in Acoustic Echo and Noise Control*. (E Hansler, G Schmidt, eds.) (Springer, New York, 2006), pp. 485–515
7) G Hu, D Wang, A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans. Audio Speech Lang. Process. **18**(8), 2067–2079 (2010)
8) G Hu, D Wang, Auditory segmentation based on onset and offset analysis. IEEE Trans. Audio Speech Lang. Process. **15**(2), 396–405 (2007)
9)G Hu, *Monaural speech organization and segregation*. (The Ohio State University, PhD thesis, 2006)
10) G Hu, D Wang, Segregation of unvoiced speech from non-speech interference. J. Acoust. Soc. Am. **124**, 1306–1319 (2008)
11) K Hu, D Wang, Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1600–1609 (2011)
12) Y Shao, S Srinivasan, Z Jin, D Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition. Comput. Speech Lang. **24**, 77–93 (2010)
13) R Meddis, et al., Simulation of auditory-neural transduction: further studies. J. Acoust. Soc. Am. **83**(3), 1056–1063 (1988)
14) D Wang, Tandem algorithm for pitch estimation and voiced speech segregation (2010). http://www.cse.ohio-state.edu/pnl/software.html, Accessed 23 September 2012
15) D Wang, G Hu, *Unvoiced speech segregation*, vol. 5. (IEEE, Toulouse, 2006), pp. 953–956
16) Y Shao, D Wang, Model-based sequential organization in cochannel speech. IEEE Trans. Audio Speech Lang. Process. **14**(1), 289–298 (2006)
17) C Rafael, E Richard, L Steven, *Digital image processing usingMATLAB (Publishing House of Electronics Industry*, (Beijing, 2009)
18) Y Lee, O Kwon, Application of shape analysis techniques for improved CASA-based speech separation. IEEE Trans. Consum. Electron. **55**(1), 146–149 (2009)
19) R Pichevar, J Rouat, A quantitative evaluation of a bio-inspired sound segregation technique for two-and three-source mixtures sounds, *Lecture*