IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

Enhancement of Speech to Text Synthesis

Ashwini V[1], Janani R[2]

U.G. Student, Department of Electronics and Communication Engineering, Meenakshi Sundararajan Engineering College, Kodambakkam, Tamil Nadu, India[1,2]

ABSTRACT:A real time speech to text conversion system converts the spoken words into text exactly as pronounced . We created a real time speech recognition system that was tested in real time noiseous environment. We used the design of a Extended Kalman filter with the dual channel microphone to enhance the ability of this Real time speech recognition system. Extended Kalman filter has been proved the best noise estimator in non-stationary noiseous environment. The purpose of this project was to introduce a new speech recognition system that is computationally simple and more robust to noise than the HMM based speech recognition system. MFCC features of speech sample were calculated and words were distinguished according to the feature matching of each sampled word. This paper deals with speech enhancement in dual-microphone devices using beam forming along with post filtering techniques. The performance of these algorithms relies on a good estimation of the acoustic channel and speech and noise statistics. In this work we present a speech enhancement system that combines the estimation of the relative transfer function (RTF) between microphones using an extended Kalman filter framework with a novel speech presence probability estimator intended to track the noise statistics' variability. The available dual-channel information is exploited to obtain more reliable estimates of clean speech statistics. Noise reduction is further improved by means of postfiltering techniques that take advantage of the speech presence estimation. Our proposal is evaluated in different reverberant and noisy environments when the system is used in both close-talk and far-talk positions. The experimental results show that our system achieves improvements in terms of noise reduction, low speech distortion and better speech intelligibility compared to other state-of-the-art approaches.

KEYWORDS:Dual-microphone; beam forming; relative transfer function; speech presence probability; post filtering; Natural language processing; Speech recognition; speech enhancement; bidirectional Kalman filter.

I. INTRODUCTION

The field of Natural language processing has always been a good research area from past years. There are numerous applications of Natural Language Processing. Speech recognition is one of the most important applications of Natural Language Processing. Speech has always been the most important part of our day to day communication. We express our ideas through a specific language. Computers understand our language (natural language) by speech recognition. Speech or word by word recognition is the process of extracting the attributes of speech and classifying the same attributes with the prerecorded datasets. To recognize a word, word must be passed on to higher-level software for syntactic and semantic analysis. It is a technique of pattern matching, where audio signals are tested and framed into phonetics (number of words, phrases and sentences). To perform such task one needs to record a voice sample and then convert this voice sample into wav format. Spectrum based parameters are obtained when a word is recognized. Various statistical methods are used for the analysis of words which give some specific value of words. Words fluctuate between in its bounded range of occurrence. In the improvement of word recognition process, one of the important tasks is to find the most informative parameters of speech signal. To perform such tasks some of the techniques are used Linear Predicted coefficients (LPC) and Mel Frequency Cepstrum Coefficients (MFCC) . By using such techniques new spectrum is obtained that is different from the previous spectrum of spoken words. Speech intensification is the process of amelioration in comprehensibility or quality of a speech sample when it depraved. Speech enhancement is not only to reduce noise from a speech sample but to de-reverberate and separate the unconstrained signals. It is desirable to enhance the speech because when speech is processed through any of the tool in lab it get influenced with the noise (background noise or otherwise) and individuality of the speech changes with time which affects the whole recognition process. So, it has become very difficult task for the boffins to find contrivances that really work in different practical environments. However this criterion plays an important role in justifying the performance of the algorithm with reference to quality and comprehensibility. Many current devices include several microphones, so that multi-channel speech processing techniques can be applied to reduce the distortions, which improves the noise reduction performance compared to single-channel approaches. This is our research focus. The most

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

common multi-channel speech processing technique is beamforming , which applies spatial filtering to the noisy speech signals captured by several microphones. One of these beamformers is the well-known Minimum Variance Distortionless Response (MVDR) beamformer, which has the advantage of being able to reduce the noise power without introducing speech distortion. The performance of MVDR depends on an accurate estimation of the noise spatial characteristics and the acoustic transfer function (ATF) between the target speaker and the microphones. For the estimation of the noise spatial statistics, methods based on multi-channel speech presence probability (SPP) are commonly used when there is no prior knowledge of the signal propagation or the spatial structure of the noise. These techniques update the noise information when speech absence is detected, which allows for the tracking of time-varying noise signals.

However, the estimation of ATFs is a more challenging task, as these depend on the speaker's location, room acoustics and microphone responses. One possible solution is based on the estimation of the beamformer weights for a certain reference microphone. As a result, the problem becomes that of estimating a relative transfer function (RTF) between the reference microphone and the rest of microphones . The most common RTF estimators are based on sub-space search, using estimates of the noisy speech and the noise spatial correlations. This category includes techniques such as covariance subtraction , covariance whitening and eigenvalue decomposition. Other approaches based on online expectation-maximization have been analyzed to jointly estimate the RTF and the clean speech and noise statistics under the assumption of a spatially-white noise field. Recently, we proposed an extended Kalman filter (eKF) framework for the estimation of the RTF between two microphones and evaluated its performance on a dual-microphone under different noisy and reverberant environments. We showed that this technique obtains better estimation accuracy than the sub-space-based estimators while allowing for the tracking of the RTF variability, especially in highly reverberant scenarios. In this paper we have used the Extended Kalman filter as RTF estimator for the intensification of our speech recognition system in background noise.

II. PROPOSED METHODOLOGIES

PROCESS AND DESIGN OF SPEECH TO TEXT CONVERSION

The process of speech recognition system is divided in two stages, first is training stage and second is testing.

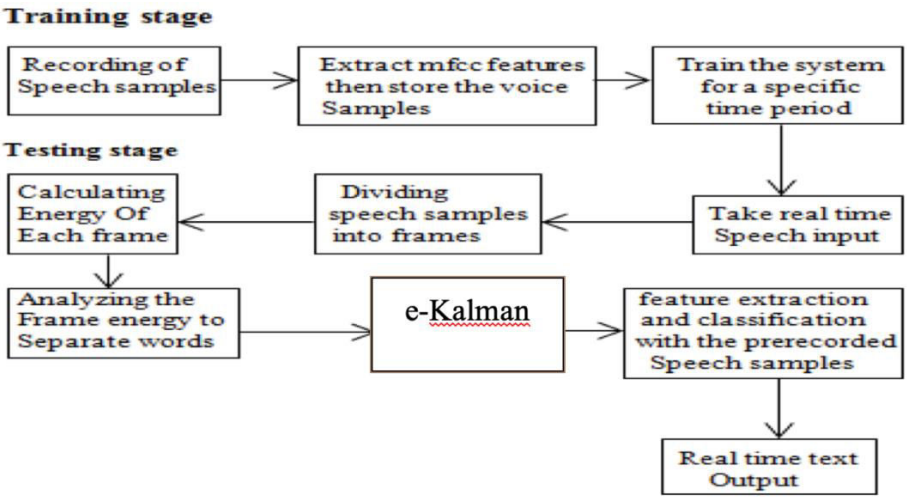

Figure 1: Speech Recognition Process

A. Feature extraction

Every signal is composed of some features/attributes. According to its features we can classify the signal characteristics. In case of speech we extract some of its attributes. Attributes extraction is one of the easiest ways to recognize the speech. Speech is a time varying signal and to deal with such a time varying signal is a difficult task. So attributes of speech play an important role in recognition. To deal with a large sequence of speech is short time attributes are taken on Mel scale (melody of speech). So we decided to extract short time features of speech which are MFCC.

We have used MFCC attributes for recognition of each word for this system. The word 'Mel' in the MFCCs represents the melody of a speech signal. MFCC features are based on the human ear perception that means human's ear's critical bandwidth frequencies filters the spaced linearity between the high frequency and low frequency of each word uttered

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

by the user. The human understanding for different frequency ranges of the uttered word is shown on a nonlinear scale. Pitch period of every word is measured with a Mel scale.

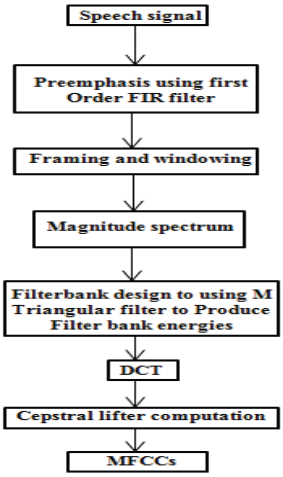

Figure 2: MFCC Feature Extraction

Experimental testing

Our speech recognition system was a speaker dependent system. So it was dependent on the user's voice only. In the testing of this system we created a database of nine words. After the training of this system, a real time speech input was given to it through a good quality dual channel microphone. The system divided the real time speech sample into small segments of frames or continuous groups of samples. After that the energy of each frame segment was calculated using simple energy formula:

$$E_x = \int_{-\infty}^{\infty} x^2\, dx \qquad (1)$$

Energy calculated was then analyzed by a speech detection algorithm to separate the words.

A. Speech disclosure algoritm

The speech disclosure algorithm is applied to detect each word by processing the stored speech samples in database (derived or self-created) frame by frame with a simple loop operation performed using Python. We divided the whole frame into the segment of 160 samples and each of the samples was detected by the system. For the detection of each frame we used a combination of signal energy and a zero crossing rate. This calculation became very simple with the python mathematical and logical operators.

B. Acoustical model

It is very important to create an acoustical model for the detection of each uttered word. So we created an acoustical model. It is known that different sounds are produced by human vocal cord and different sounds can have different frequencies. To predict the different frequencies it power spectral density measure can be a better way. So we find out the frequencies by power spectral densities measure. Speech can be termed as short term stationary so MFCC features were again extracted and word pronounced by the user was detected. The output speech signal was compared with the prerecorded clean speech signal with the correlation figure where we had taken the three voice samples for each of the word training. Correlation figure was calculated using :

$$\rho(w_0, w_1) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{oij} * w_{rij}}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{oij}^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{rij}^2}} \qquad (2)$$

C. Applying extended kalman filter algorithm for noise reduction using dual channel microphone

The proposed enhancement system for dual-microphone smartphones is depicted in Figure 3

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

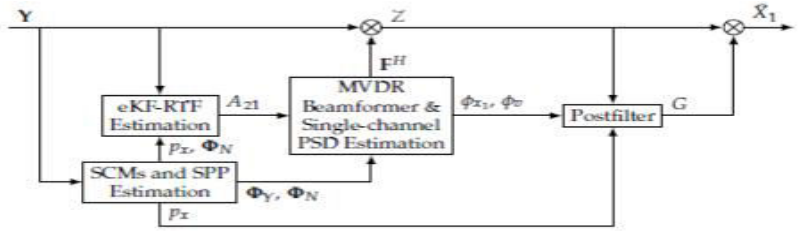

Figure 3. Overview of the dual-microphone enhancement system

The microphones capture the noisy speech signals $y_m(t)$, where m indicates the microphone index (m = 1, 2). We assume an additive noise distortion model which in the short-time Fourier transform (STFT) domain can be expressed as

$$Ym(t,f) = Xm(t,f) + Nm(t,f) \quad (3)$$

where $Ym(t,f), Xm(t,f)$ and $Nm(t,f)$ represent, respectively, noisy speech, clean speech and noise signal STFTs, t is the frame index and f the frequency bin. The two channel components can be stacked in a vector,

$$Y(t,f) = [Y1(t,f) \quad , \quad Y2(t,f)]^T \quad (4)$$

where $[.]^T$ indicates matrix transposition. Similarly, we define vectors $X(t,f)$ and $N(t,f)$. In the following, we will consider that each frequency component can be processed independently from the others, which is commonly referred to as narrowband approximation.

As shown in Figure 3, the beamforming block requires two previous procedures. First, a speech presence probability (SPP)-based algorithm is employed for the estimation of the noisy speech and noise spatial correlation matrices (SCM), $\phi_Y(t,f)$ and $\phi_N(t,f)$, respectively, and the SPP, $px(t,f)$.

Then, an extended Kalman filter (eKF)-based estimator obtains the relative transfer function (RTF) between the secondary microphone (m = 2) and the reference microphone (m = 1), defined as

$$A21(f,t) = \frac{X2(f,t)}{X1(f,t)} \quad (5)$$

Once the SCM matrices and RTF have been computed, the dual-channel noisy speech vector $Y(t,f)$ is processed using a Minimum Variance Distortionless Response (MVDR) beamformer, which applies spatial filtering yielding an output signal $Z(t,f)$ defined as

$$Z(t,f) = F^H(t,f)Y(t,f) \quad (6)$$

where $(.)^H$ indicates a Hermitian transpose. The coefficients of the beamformer, F(t, f), are estimated as [2],

$$F(t,f) = \frac{\phi_N^{-1}(t,f)\, d(t,f)}{d^H(t,f)\phi_N^{-1}(t,f)d(t,f)} \quad (7)$$

Where,

$$d(t,f) = [1 \quad , \quad A_{21}(t,f)]^T \quad (8)$$

is the steering vector normalized to the reference channel.

Finally, the speech signal at the beamformer output is enhanced by a single-channel postfilter for additional noise reduction. A spectral gain $G(t,f)$ is obtained using the power spectral density (PSD) of the speech at the reference microphone $\phi_{x_1}(t,f)$, the PSD of the residual noise $\phi_v(t,f)$ and the SPP $p_x(t,f)$. The above single-channel statistics ($\phi_v(t,f)$ and $\phi_{x_1}(t,f)$) are estimated from the noisy speech and noise SCMs and the RTF. The gain function is further processed by a musical noise reduction algorithm and finally applied to the beamformer output, thus obtaining

$$\widehat{X_1}(t,f) = G(t,f)Z(t,f) \quad (9)$$

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

where $\widehat{X_1}(t,f)$ is the estimate of the clean speech signal at the reference microphone. In the following sections, the description of the different parts of the system is addressed.

III.EXTENDED KALMAN FILTER BASED RELATIVE TRANSFER FUNCTION ESTIMATION

The proposed system requires knowledge of the RTF between the two microphones, namely $A_{21}(f,t)$. Among other approaches, the computation of this RTF can be addressed as the estimation of a variable that changes across frames in terms of a stochastic model. Given this dynamic model, the noise statistics and the noisy observations, the tracking of the RTF using an extended Kalman filter (eKF), showing a better estimation performance in comparison with other state-of-the-art approaches.
Kalman gain,

$$\mu_{2,t} = E[y_{2,t}]$$
$$S_{y,t} = E[(y_{2,t} - \mu_{y,t})(y_{2,t} - \mu_{y,t})^T]$$
$$C_{ay,t} = E[(a_{21,t} - \hat{a}_{21,t|t-1})(y_{2,t} - \mu_{y,t})^T]$$

To deal with non-linear functions, two variants of the Kalman filter are widely used: the eKF and the unscented Kalman filter (uKF) . In the case of the model of , it could be demonstrated that both eKF and uKF yield the same closed-form expressions for $\mu_{y,t}, S_{y,t}$ and $C_{ay,t}$. We choose the eKF approach because it gives a more stable computational solution and it is easier to implement than the uKF one. In the next subsection the eKF approach is presented.

Postfiltering Approaches for Dual-Microphone
As introduced above, the performance achieved by beamforming algorithms is limited in our scenario mainly due to the reduced number of microphones. In this section we propose different alternatives for single-channel postfiltering at the beamformer output. These postfilters make use of the previously estimated parameters to design again function which is applied to the beamformer output signal for further noise reduction. In the next subsections we describe several postfiltering techniques as well as our proposals for the estimation of the clean speech and noise single-channel statistics required by the former ones.

Parametric Wiener Filtering
Under the distortionless constraint of MVDR, if we assume an accurate estimation of the RTF, we can approximate the single-channel speech signal *Z(t, f)* at the beamformer output as

$$Z(t,f) \simeq X_1(t,f) + V(t,f),$$

where *V(t, f)* is the residual noise at the beamformer output,

$$V(t,f) = F^H(t,f)N(t,f)$$

The PSDs of the clean speech signal and the residual noise *V(t, f)* are given by $f_{xl}(t,f)$ and $f_v(t,f)$, respectively. The corresponding Wiener filter (WF) for $X_l(t,f)$ is then defined as

$$G(t,f) = \frac{\varepsilon(t,f)}{1+\varepsilon(t,f)}$$

where

$$\varepsilon(t,f) = \frac{\varphi_{x_1}(t,f)}{\varphi_v(t,f)}$$

is the a priori SNR.
The noise reduction performance of WF can be improved if we consider the a posteriori SPP $p_x(t,f)$ in the postfiltering design. Our goal is to decrease the gain factor for time-frequency bins where speech is absent. This yields the parametric WF (pWF) [34]

$$G(t,f) = \frac{\varepsilon(t,f)}{\mu(t,f)+\varepsilon(t,f)}$$

where *m(t, f)* is an SPP-driven trade-off parameter. To obtain this parameter, we use the same mapping function as ,

$$\mu(t,f) = \mu_{\min} + \frac{(\mu_{max}-\mu_{min})10^{\frac{c\rho}{10}}}{10^{\frac{c\rho}{10}}+\left(\frac{p_x(t,f)}{1-p_x(t,f)}\right)^{\rho'}}$$

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

which is similar to the mapping function of . We can see that $\mu(t,f) \in [\mu_{min}, \mu_{max}]$ and therefore, the lower the a posteriori SPP, the higher $\mu(t,f)$, thus improving the noise reduction performance.

Optimally Modified Log-Spectral Amplitude Estimator

Instead of estimating a WF filter over the signal statistics, other approaches try to directly make an MMSE estimation of the clean speech signal amplitude $|X_1(t,f)|$. One of these is the log-spectral amplitude (LSA) estimator and defined as

$$\begin{aligned}\left|\hat{X}_1(t,f)\right| &= exp\,(E\,[\,log\,(|X_1(t,f)|)|\,Z(t,f)|]) \\ &= G_{H_x}(t,f)\,|Z(t,f)|\end{aligned}$$

where

$$G_{H_x}(t,f) = \frac{\varepsilon(t,f)}{1+\varepsilon(t,f)} \exp\left(\frac{1}{2}\int_{\frac{\varepsilon(t,f)}{1+\varepsilon(t,f)}\gamma(t,f)}^{\infty} \frac{e^{-u}}{u} du\right)$$

is the LSA gain function, and

$$\gamma(t,f) = \frac{|Z(t,f)|^2}{\phi_v(t,f)}$$

is the a posteriori SNR. As can be seen, this estimator is equivalent to applying a gain function over the noisy speech signal $Z(t,f)$ that depends not only on the a priori SNR but also on the a posteriori SNR. The obtained gain function is similar to WF for high a priori SNRs and has lower values than WF when the a priori SNR is lower and the a posteriori SNR is higher (i.e., when noise tends to dominate).

The estimator of (70) is only valid under the assumption of speech presence. To improve the performance, the authors proposed a different estimator that also takes into account the a posteriori SPP and different gain functions for speech presence and absence. That is the optimally modified LSA (OMLSA) estimator, which is defined as

$$G(t,f) = G_{H_x}(t,f)^{p_x(t,f)} G_{H_v}(t,f)^{1-p_x(t,f)}$$

where G_{Hv} is a constant attenuation applied when speech is absent, whose value is usually -25dB .

5.3. Single-Channel Speech and Noise PSD Estimators

The computation of the gain functions previously presented relies on the estimation of the single-channel PSDs of the clean speech at the reference channel $f_{x1}(t,f)$ and the residual noise at the beamformer output $f_v(t,f)$. The estimation of the residual noise PSD is straightforward if estimates of the SCM $\boldsymbol{\phi}_N(t,f)$ and the steering vector $d(t,f)$ are available. In this case, the residual noise PSD can be obtained at each time-frequency bin as

$$\hat{\phi}_v = \left(\hat{d}^H \hat{\phi}_N^{-1} \hat{d}\right)^{-1}$$

On the other hand, a clean speech PSD estimate is more difficult to obtain because of its higher variability. We proposed two different estimators that make use of the noisy speech and noise statistics and the RTF between microphones. In the following, we describe both clean speech PSD estimators, omitting time and frequency indices for simplicity.

5.3.1. Power Level Difference-Based Estimation

The PLD-based estimator is derived from the method , which exploits the PLD between the microphones of a smartphone used in CT conditions; that is, a more attenuated clean speech component is expected at the secondary microphone with respect to the primary one. This PLD is defined in terms of the noisy speech PSDs at the microphone inputs

$$\Delta\hat{\varphi}_{PLD} = \max(\hat{\varphi}_{Y_{11}} - \hat{\varphi}_{Y_{22}}, 0),$$

where it is assumed that the power at the reference microphone is always higher than the one at the secondary microphone. Assuming that the noise PSD difference $\left(\Delta\varphi_N = \varphi_{N_{11}} - \varphi_{N_{22}}\right)$ can be neglected when compared to $\Delta\hat{\varphi}_{PLD}$ the clean speech PSD can be estimated as in [17],

$$\hat{\varphi}_{x_1} = \frac{\Delta\hat{\varphi}_{PLD}}{1-|\hat{A}_{21}|^2}$$

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

Although this estimator offers good performance in CT conditions, the previous assumptions are not longer valid in FT conditions.

Minimum Variance Distortionless Response-Based Estimation

This estimator calculates the PSD directly at the beamformer output by spectral subtraction,taking into account the distortionless property of MVDR. The estimator is defined as

$$\hat{\varphi}_{x_1} = F^H(\hat{\varphi}_Y - \hat{\varphi}_N)F$$

so that it fully exploits the spatial information by using the SCM matrices of the noisy speech and noise signals. The combination of both channels by means of the beamformer weights (F) allows for a more robust estimation than directly taking the first element of $\hat{\varphi}_Y - \hat{\varphi}_N$. Negative PSDs may be obtained due to the estimator variance, is bounded by 0.

IV. EXPERIMENTAL RESULTS

The performance of the different estimators and speech enhancement algorithms discussed along this paper is evaluated by means of objective speech quality and intelligibility metrics. Two different well-known objective metrics are used:

- The Perceptual Evaluation Speech Quality (PESQ) metric is utilized to evaluate the speech quality of the enhanced speech signal. This metric gives a mean opinion score between one and five. The higher the PESQ values, the better the speech quality.
- The Short-Time Objective Intelligibility (STOI) metric is intended to evaluate the speech intelligibility of the enhanced speech signal. The resulting score is a value between zero and one.The higher the STOI value, the better the speech intelligibility.

PESQ and STOI are both intrusive metrics, which means that they need a clean reference. As a reference, we use the clean speech signal at the reference microphone, $x_1(n)$. Additionally, in order to evaluate the RTF estimation accuracy, we use the speech distortion (SD) index . This index measures the distortion level on the clean speech signal at the beamformer output, namely $\tilde{x}_1(n)$(inverse STFT of $F^H(t,f)X(t,f)$). The idea behind using this metric is that, because of the distortionless property of MVDR, more accurate RTF estimates should yield lower speech distortion.
The SD index is measured segmentally across the speech signal, in such a way that the SD value at the i-th segment is obtained as

$$SD(i) = \frac{\sum_{n=(i-1)T}^{iT-1}|x_1(n)-\tilde{x}_1(n)|^2}{\sum_{n=(i-1)T}^{iT-1}|x_1(n)|^2}$$

where T is the number of samples per segment. The segmental SD values are averaged to achieve the final SD index, which is a value between zero and one. The lower the SD value, the lower the speech distortion. As silence frames are excluded from this evaluation by means of calculating the median of the segment-wise signal power and removing those segments with a power 15 dB lower than that median. We evaluate the proposed algorithms by using simulated dual-channel noisy speech recordings from a dual-microphone smartphone. Two different databases were developed for each device use mode: close-talk (CT) and far-talk (FT). To simulate the recordings, clean speech signals are filtered using dual-channel acoustic impulse responses and real dual-channel environmental noise is added at different signal-to-noise ratios (SNRs). We evaluate four different reverberation environments and eight different noises, which are matched as indicated in Table 1, yielding eight different acoustic environments (including both reverberation and noise).
Table 1. Predefined acoustic environments: each environment combines a reverberation environment with a given noise.

Reverberation	Noise (Test Only)
(A) No reverb	Car, Street, Pedestrian street
(B) Low	Bus, Café
(C) Medium	Babble, Bus station
(D) High	Mall

Clean speech signals are obtained from the TIMIT database and downsampled to 16 kHz.In particular, a total of 850 clean speech utterances from different speakers are employed. All the utterances have a length of around seven seconds, which is achieved by same-speaker utterance concatenation. Two different sets are then defined, namely training and test. Speakers do not overlap across sets. Moreover, the number of utterances from female and male speakers is balanced. The distribution of utterances and the number of speakers in the training and test sets are indicated in Table 2.
Table 2. Distribution of clean speech utterances and speakers across training and test sets.

SN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006401 |

	N° Utterances	N° Speakers
Training set	700	440
Test set	150	93

The training set consists of reverberated clean speech utterances and is only used to estimate the a priori statistics for the eKF-RTF estimator. Such utterances were obtained by filtering each clean speech utterance with a set of dual-channel acoustic impulse responses which model four reverberant environments, thereby yielding a total of 2800 training utterances. Sixteen different acoustic impulse responses were considered for each acoustic environment. For each utterance, the impulse response was randomly selected. On the other hand, the test set consists of utterances contaminated according to the eight noisy environments defined in Table 1. This set is intended to evaluate the different algorithms proposed in this work. Noises were added to the reverberated speech at six different SNRs from -5 dB to 20 dB, so a total of 7200 test utterances was obtained. In order to simulate reverberation, ten different acoustic impulse responses, in turn also different from those of the training set, were considered for each acoustic environment and, again, randomly applied. For STFT computation, a 512-point DFT was applied using a 32 ms square-root Hann window with 50% overlap. This results in a total of 257 frequency bins for each time frame. The noisy speech and noise SCMs were estimated using an updating constant ˜a = 0.9. The parameters for CDR-based a priori SAP calculation were set as [5]: d_{mic} = 13 cm, q_{min} = 0.1, q_{max} = 0.998, c = 3 and ρ= 2.5. The SPP threshold for RTF updating was set as p_{thr}= 0.9. Finally, for the SPP-driven trade-off parameter of the parametric Wiener filter, the following parameter values were used [5]: μ_{min}= 1,μ_{max} = 4, c=-3 and ρ = 4.
The algorithm was implemented in Python . The computational burden of the implementation was evaluated on a PC with an Intel Core i7-4790 CPU. The algorithm works on a frame-by-frame basis, so that the algorithmic delay is in this case the duration of a frame, that is, 32 ms. The average performance of the whole system (i.e., including SPP and RTF estimation, MVDR beamforming and postfiltering) on this machine achieved approximately 8x faster than real-time.

In the following, the performance of the different techniques is tested. In order to simplify the reading of the results tables, an acronym is provided (in parentheses) for each considered technique after its description

V. CONCLUSION

We have developed speech recognition system and used a dual channel microphone with extended Kalman filter for its enhancement. We created our own database in Python and also used TIDIGIT dataset for its comparison analysis. We concluded that in comparison with the TIDIGIT dataset this system yields excellent results in real time scenarios while with our own created database the results are comparatively poor. Only five sentences and five similar words are shown in this paper. We tested this system in a real time environment for different noise conditions and compared the output with the prerecorded speech samples with the correlation figure. First we tested this system at home with fan noise. System was able to recognize each word and gave us 100% accuracy in that case. After that we tested this system with fan noise and background music. Based on this analysis our system is overall 90% accurate. This system's accuracy can be increased in more different noise scenarios.
By using this code the system can be trained for more words and paragraphs. Kalman filter removes background noise very efficiently. But this filter takes large time to filter out noise. In case or continuous acquisition of speech it can take more time to display text because of its word by word filtration process.

REFERENCES

[1] Parchami, M.; Zhu,W.P.; Champagne, B.; Plourde, E. Recent developments in speech enhancement in the short-time Fourier transform domain. IEEE Circuits Syst. Mag. 2016, 16, 45–77. [CrossRef]
[2] 2. Benesty, J.; Chen, J.; Huang, Y. Microphone Array Signal Processing; Springer: Berlin, Germany, 2008; Volume 1.
[3] 3. Kumatani, K.; McDonough, J.; Raj, B. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. IEEE Signal Process. Mag. 2012, 29, 127–140. [CrossRef]
[4] 4. Souden, M.; Benesty, J.; Affes, S.; Chen, J. An Integrated Solution for Online Multichannel Noise Tracking and Reduction. IEEE Trans. Audio Speech Lang. Process. 2011, 19, 2159–2169. [CrossRef]
[5] 5. Taseska, M.; Habets, E. Nonstationary noise PSD matrix estimation for multichannel blind speech extraction. IEEE/ACM Trans. Audio Speech Lang. Process. 2017, 25, 2223–2236.
[6] Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. IEEE/ACM Trans. Audio Speech Lang. Process. 2017, 25, 692–730. [CrossRef]

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details