SPEECH CLASSIFICATION FOR ENHANCING SINGLE CHANNEL BLIND DEREVERBERATION

Steven A. Fortune and James R. Hopgood

Institute for Digital Communications, University of Edinburgh,
The King's Buildings, Mayfield Road, Edinburgh, United Kingdom, EH9 3JL
phone: + (44) (0)131 650 5565, fax: + (44) (0)131 650 6554, email: {steven.fortune, james.hopgood}@ed.ac.uk

ABSTRACT

Several single channel dereverberation techniques exists that enhance the harmonic properties of voiced speech, or utilise a signal model of unvoiced speech. This paper demonstrates how existing speech dereverberation methods can be improved by classifying speech into voiced, unvoiced and silent segments. Methods that enhance the harmonic features of voiced speech can benefit from enhancing only voiced segments, compared to using the entire signal. Removing silent frames from the input can additionally benefit dereverberation methods. Additional features that can be used for dereverberation, signal entropy and minimising the energy of silent periods, are introduced and show good performance. However, speech classification is more difficult for reverberant speech than clean speech. The performance of a number of different classification measures are compared in a reverberant environment. It is shown how performance degrades with increasing reverberation, but some classifiers do hold their performance better than others. The accuracy of the estimation of a dereverberation filter parameter using various signal features are compared. In addition, several signal features can be combined into one cost function. This shows promise in giving improved overall estimation accuracy, by looking at and enhancing a richer set of speech features.

1. INTRODUCTION

Reverberation occurs when an audio source and receiver are separated in an enclosed space. Reverberation can reduce speech intelligibility, or reduce the accuracy of speech or speaker recognition techniques [1]. A multi-channel approach using multiple microphones can assist this problem, but is not always practical, for example in a hearing aid application. A method for single-channel dereverberation is thus desirable. The solution to this problem is intractable without the addition of prior information, such as the statistics of the source or the channel, or enhancing a known feature of speech signals degraded by reverberation.

One way to add this prior information is to devise a model for the source and the channel, and attempt to fit the observed data to the models. A solution may be found if the channel is stationary and the source varies sufficiently [2]. This allows separation of channel and source, and thus equalisation of the channel [3]. However, this method is sensitive to errors in the channel estimate when it comes to inverse filtering. It also does not take account of the rich harmonic information available in voiced speech as the speech model assumes unvoiced speech.

An alternative approach is to measure and enhance a certain feature of the speech that is degraded by reverberation. This feature can then be used to drive an adaptive filter maximising the feature in the estimated speech as shown in Fig. 1. Possible features used for dereverberation include harmonicity of voiced speech [4] and kurtosis of the linear prediction (LP) residual [5]. This feature based approach can also be extended into a probabilistic framework [6] using statistical source models [7]. This approach would allow the filter to adapt to a slowly varying channel.

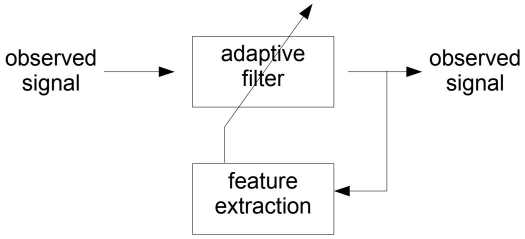

Figure 1: *Feature extraction method for speech dereverberation*

By using the *feature extraction* framework in Fig. 1, different signal features can be compared directly, something that has not been performed before. In addition, multiple features could be combined into a single *cost function*. This could deliver improved accuracy as more information can be included from the signal. However, a combined cost function is likely to require a computationally expensive optimisation routine. Using a single feature it is often possible to derive a faster filter adaption method using the gradient [5].

The basis behind any feature extraction technique is that speech has a certain structure, be it harmonicity or a certain pdf, and that this structure is reduced by reverberation. If a measure of this structure can be made, the measured feature could potentially be maximised to remove the effect of the dereverberation. Many blind dereverberation techniques treat all speech the same. However, the structure of speech changes rapidly, with each sound having different possible excitations, white noise, glottal pulse or none, giving rise to unvoiced speech, voiced speech and silence respectively. It is of the authors opinion that the structure of these three classifications of speech are different enough that they should be treated differently by a feature extraction method. For example, by enhancing the harmonicity of voiced speech only, not the entire signal, or by removing silent frames from the estimation input.

In this paper, it is proposed that a stage of speech classification, into voiced speech, unvoiced speech and silence, is performed as part of the feature extraction. Performing this classification is more difficult in a reverberant environment. The performance of different methods of classification on a reverberant signal is examined in section 2.3. Speech classification enables the measurement of extra features indicating the level of reverberation in the signal, and improved use of other features as discussed in Section 3. The performance of these features in estimating the inverse filter is tested in Section 4.

2. SPEECH CLASSIFICATION

2.1 Speech Model

Speech sounds can be divided into three broad classes depending on this mode of excitation [8, 9]. **Voiced sounds** such as *aah* are produced by vibrating vocal cords producing a periodic series of

IDCOM is in the Joint Research Insititue for Signal and Image Processing, a member of the Edinburgh Research Partnership in Engineering and Mathematics. This work was supported by EPSRC (UK research council) grant EP/D051207/1

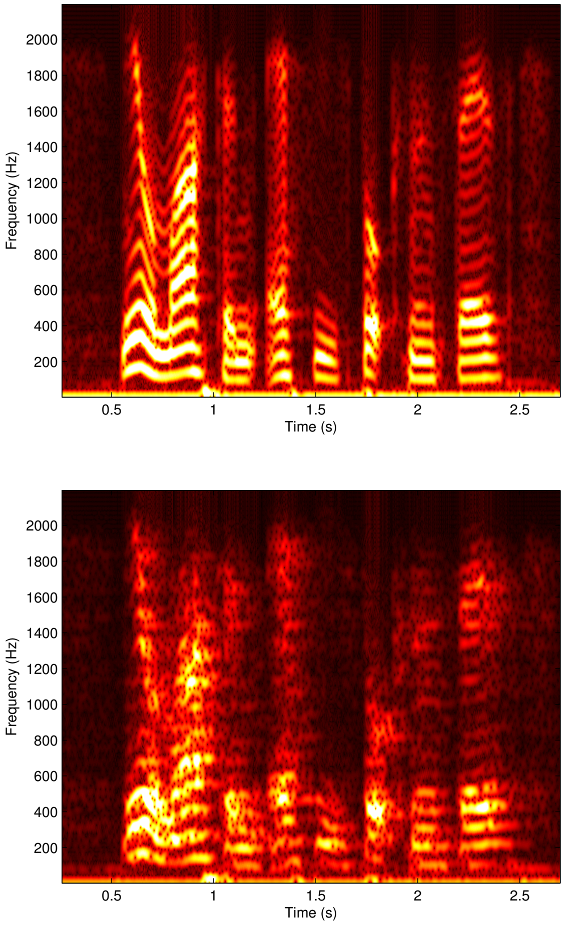

Figure 2: *Spectrogram of original speech segment and reverberated speech segment with reverberation time of 0.25s.*

glottal pulses. The sound is quasi-periodic with a spectrum of rich harmonics at multiples of the fundamental or pitch frequency f_0 as shown in Fig. 2. **Unvoiced sounds** do not have a vibrating source. They are produced by turbulent flow, leading to a wide-band noise source. Plosive sounds, with an impulsive source, also exist. They are generally of very short duration, making them less important, and can be adequately described the the unvoiced model. In the course of speech it is also normal to have periods of silence, when no excitation is present.

2.2 Speech Classification

A speech/silence classifier is often referred to as a voice activity detector (VAD) and can be combined with a voiced/unvoiced classifier. A number of different classification methods are examined in Section 2.2.1, and compared in a reverberant environment in Section 2.3.

2.2.1 Classification Parameters

Speech is generally divided into short (10-30ms) frames either adjacent, or overlapping. A parameter is extracted from each frame, and a decision on classification made based on the parameter value. A large number of different parameters can be used for speech classification. A brief overview follows.

- **Energy** of a frame can be a basic measure of whether speech is present. A simple fixed threshold can be used, or preferably an adaptive threshold, which measures an adapts to a changing noise floor [10, 11].
- **Zero Crossing Rate** or ZCR can give an indication of the spectral properties of a signal [9]. Rabiner and Sambur [10] showed the ZCR combined with a short term energy measurement can be used for VAD. Furthermore, the distribution of ZCR is different between voiced (5-20 per 10ms) and unvoiced (30-70 per 10ms) speech [9]. This indicates ZCR could be used to discriminate between different classes of speech.
- **Periodicity** A way to discriminate between voiced and unvoiced sounds is the proportion of the signal which is periodic. A least-squares periodicity estimator (LSPE) using the time domain signal $s[n]$ for $n = 1, \cdots, N$, varies the estimated pitch period, $\hat{P}_0$, to maximise the periodicity given by [11]

$$\begin{aligned} R(\hat{P}_0) &= \frac{I_0(\hat{P}_0) - I_1(\hat{P}_0)}{\sum_{n=1}^{N} s[n]^2 - I_1(\hat{P}_0)}, \\ I_1(\hat{P}_0) &= \sum_{n=1}^{\hat{P}_0} \sum_{h=0}^{K_0} s[n + h\hat{P}_0]^2 / K_0, \\ I_0(\hat{P}_0) &= \sum_{n=1}^{\hat{P}_0} \left(\sum_{h=0}^{K_0} s[n + h\hat{P}_0] \right)^2 / K_0, \end{aligned} \tag{1}$$

  where $K_0 = (N - n)/\hat{P}_0 + 1$. The periodicity measure [12, 11] gives an indication of the proportion of periodic component in the signal. A fixed threshold can be used to detect voiced frames [11].
- **Harmonic Frequency Domain Analysis Based on a Sinusoidal Speech Model**
  The spectrogram of a voiced signal (Fig. 2) shows a peak not only at the pitch frequency $f_0 = 1/P_0$, but also at the harmonics, $2f_0$, $3f_0 \cdots$. A more robust technique would look for these harmonic peaks in the spectrum. This can be performed in the frequency domain [8]. First the pitch can be estimated by fitting a harmonic set of sinusoids to the input data [8, 13]. The vocal tract envelope, $\bar{V}(f)$, can be estimated using the Spectral Envelope Estimation Vocoder (SEEVOC) [14]. The degree to which the data fits the harmonic model can be used to determine the degree of voicing in the signal [13].

2.3 Performance on reverberant speech

In this section, the performance of different parameters described in Section 2.2.1, when combined with an energy measurement, are evaluated. A section of speech was classified manually and the accuracy of different classification techniques compared to this. The performance was evaluated on the original speech signal, then the signal filtered by a synthetic channel of increasing reverberation times. The channel response was generated using white Gaussian noise with a negative exponential envelope following a method by Habets [15]. Thus the channel impulse response $h[t]$ is given by

$$h[t] = \begin{cases} b[t]e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases},$$

where $b[t]$ is zero mean white Gaussian noise and $\alpha = 3\log(10)/T_r$ and T_r is the reverberation time. The speech segment used was 10s long, with 14% silence, 59% voiced speech, 23% unvoiced speech, 2% plosives and 2% unknown.

The results are shown in Fig. 3. With no reverberation, a periodicity classifier performs better than one using the ZCR. However the ZCR is affected less by reverberation and outperforms periodicity with moderate levels of reverberation. This could be explained by examining the periodicity values measured throughout the signal. For voiced speech, reverberation reduced the periodicity, but for unvoiced speech, the periodicity measure increased. This is caused by any reverberant peaks in the channel, plus any harmonic components time-shifted into those frames. The change in ZCR levels due to reverberation was less significant. The harmonic sinusoidal model method produced the best results, and shows the highest resistance to reverberation. This is likely due to the measure requiring regular harmonic peaks, an unlikely scenario for a reverberant channel or smeared and shifted peaks.

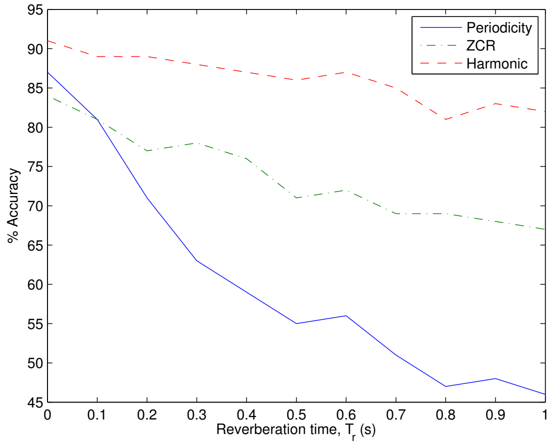

Figure 3: *Percentage accuracy of classification methods compared to manually classified data. Measurements are combined with energy measurement and performed on 10s speech segment. Speech segment is filtered with synthetic room response of increasing reverberation time.*

A number of alternate methods are available for VAD and classification. These results show that not all techniques work equally well on reverberant speech, so it is useful to have a more thorough examination of speech classification in a reverberant environment.

3. FEATURES USED FOR DEREVERBERATION

This section describes how speech classification can be used to aid dereverberation. Once the reverberant speech is classified, each speech frame can be measured with an appropriate feature or dereverberation method. Each classification will be dealt with separately below. However, using the model shown in Fig. 1 it is possible to combine several reverberant measures into a single *cost function* that could be used to optimise the inverse filter estimate as discussed in Section 3.5.

3.1 Silence

A feature of clean speech is that it contains periods of silence in between individual utterances. In a reverberant environment, these silence are filled in or smeared, as shown in Fig. 2. Thus it may be possible to estimate the dereverberation filter by minimising the energy in these silent periods. It was found that signal variance worked better than energy, due to any dc bias of individual frames. The measure C_{SIL} was calculated as the variance of the concatenation of all silent frames.

Another use of detecting periods of silence is that silent frames would give erroneous results in some dereverberation methods which assume excitation is present. By removing these frames from the estimation process an improvement may be obtained. This idea was tested using the model Bayesian approach by Hopgood et. al. [3]. A 72-order all pole channel model was used, with a 10s speech segment as input. The channel was estimated with this speech, then repeated with any silent periods removed from the input. Results are shown in Fig. 4. Removing the frames deemed as silent (13% of signal), reduced the mean squared error of the channel magnitude estimate by 10%.

3.2 Voiced Speech

Voiced speech is quasi-periodic, with a spectrum of rich harmonics at multiples of the fundamental frequency. Reverberation smears these harmonics, reducing the periodicity or harmonicity of the speech. By measuring, and maximising this harmonic structure, it is possible to estimate the dereverberation filter providing clean speech. Several measures of this harmonic structure are possible, as discussed below.

3.2.1 LP Residual Kurtosis

Linear prediction (LP) residuals of voiced speech have strong peaks corresponding to glottal pulses. A measure of amplitude spread of LP residuals, such as kurtosis, can serve as a reverberation metric [5]. To make this measurement, the signal is broken into 30 ms frames $s_m[n]$ and a 12-th order linear prediction estimate of the signal made. The kurtosis of the LP residual, or error in the estimate, is measured. Clean speech has a higher LP kurtosis than reverberated speech, so maximising the measure C_{LPK} can be used to estimate the dereverberation filter.

3.2.2 Periodicity

Periodicity (1) measures how strong the periodic component in a frame is. Reverberation reduces the periodicity of voiced speech, and increases periodicity of unvoiced speech. The dereverberation filter is estimated by maximising the measure C_{PER} which is the mean of the periodicity measurement of voiced frames only. Another measure tested was the variance of the periodicity of all speech frames C_{VPER}.

3.2.3 Harmonic Energy

In HERB [4], an adaptive harmonic filter is applied to the signal. This extracts frequency component corresponding to multiples of the given fundamental frequency f_0. To fit within the feature enhancement framework, we will make a crude measurement of harmonicity, corresponding to a measure of the energy of the signal harmonics. This is calculated by zero padding the signal frame $s_m[n]$, taking the Fourier transform giving $S_m[f]$ and summing the magnitude of the spectrum at the first K harmonics, where K is fixed to fit within the signal bandwidth, using

$$C_H[m] = \frac{1}{K}\sum_{k=1}^{K} S_m[kf_0]$$

The index of the fundamental frequency bin f_0 is chosen to maximise this energy measure for each individual frame. The total harmonic measure C_H is then the mean of $C_H[m]$ over all voiced frames.

3.3 Unvoiced Speech

Unvoiced speech can be modelled as white noise filtered by a time varying AR process, representing the vocal tract [9]. The reverberant channel may be modelled as an all-pole filter with the same structure as an AR process [3]. Thus there is no special feature of the speech that could be extracted from a single frame to aid dereverberation except for statistical differences between source and channel filters. If one considers that the source filter changes rapidly (between phonemes) and the channel filter varies slowly, this can be used to separate the source and channel models. This has been demonstrated for a stationary channel [3] and a restricted time varying channel [16].

Figure 4 shows the results of this Bayesian channel estimation for a stationary channel. A reasonable channel estimate was obtained for unvoiced frames only, even though they consisted of only 23% of the input speech segment.

3.4 Entropy

A technique used to correct for unknown phase distortions in astronomical and Synthetic Aperture Radar images is image sharpening [17]. This is a similar approach to feature extraction, where *signal sharpness* is maximised for a clean signal, and reduced by reverberation. A good measure of image (or signal) sharpness is negative

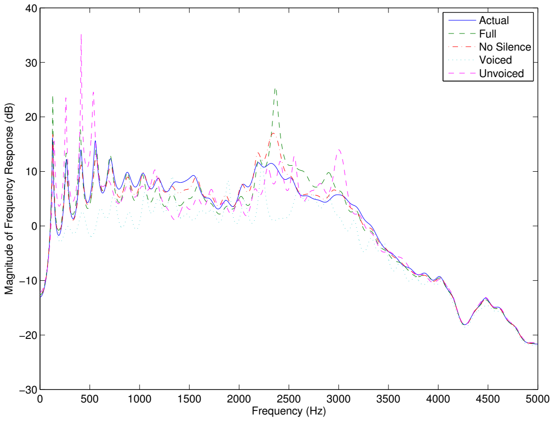

Figure 4: *Estimated frequency response of stationary channel using Bayesian method [3]. Input included all speech frames, silent frames removed, voiced frames only and unvoiced frames only. The mean squared error of magnitude estimate is 0.0031, 0.0028, 0.0035 and 0.0081 respectively. Speech segment consisted of 13% silence, 58% voiced speech, 23% unvoiced speech.*

entropy. In information theory, entropy can quantify the information contained in a signal. This information is reduced by reverberation. Thus maximising the entropy in a signal may remove the reverberation in speech. The entropy measure C_{ENT} of signal $s[n]$ is estimated using

$$C_{ENT} = -\sum_{n} |s[n]| \log |s[n]|.$$

3.5 Combined Cost

A combined cost function is calculated, which is a combination of several signal features. In this case, entropy, silence energy, LP residual kurtosis of voiced signal, mean periodicity of voiced signal and harmonic energy of voiced signal. Each measure was scaled to the interval $[0,1]$ over the parameter range measured, then summed. If an optimisation method is used, correct scaling would need to be determined prior to the test.

4. DEREVERBERATION ACCURACY

In this section, the accuracy of estimating a dereverberation filter using the various features discussed in Section 3 are analysed. The reverberant channel is modelled using a 300-th order all-pole IIR filter b_p, for $p \in \{1, \cdots, 300\}$. The corresponding dereverberation inverse filter is thus a FIR filter of the same order. The classification was performed once, on the initial reverberant speech. This was to save time, as these trials take a long time to compute. Repeating the classification every step may alter results, but we do not believe significantly so.

4.1 Single parameter tests

A single parameter of the dereverberation filter was varied, and the various features measured using the estimated signal. The parameter value that maximises each feature is determined. This is repeated for a number of different parameters and parameter values. The parameter estimates can be compared to the known channel parameter values b_p to determine the accuracy of the estimation technique. This test was performed repeatedly on a single 10s speech segment for each of the 300 filter parameters ($p \in \{1, \cdots, 300\}$), then for 100 varied values of a single parameter $b_{200} \in \{-0.5, -0.49, \cdots, 0.5\}$. The mean squared error between estimated and known parameters was calculated over all tests. The results are shown in Table 1.

	ENT	SIL	LPK	MPer	VPer	HAR	Com
Voiced	7.2	1.6	17	0.11	6.2	36	0.22
All			442	642	568	68	

Table 1: *Mean squared error in filter parameter estimates, all $\times 10^{-4}$. Features used for parameter estimation are described in text and include; entropy, silence energy, LP residual kurtosis, mean periodicity, variance of periodicity, harmonic energy and a combined measure respectively. Voiced speech measures (LPK, MPer, VPer and HAR) were measured both for all speech frames and voiced frames only.*

There is a clear performance gain in measuring the voiced features (LPK, periodicity and HAR) for voiced only segments, compared to all segments. This shows the classification step can increase the performance of dereverberation techniques that use these measures.

The new measures tested, namely entropy and silence energy, show they can successfully estimate the individual filter parameters in the tests performed. They in fact have better performance than using LPK and HAR and are easier to calculate. The mean value of periodicity of voiced frames gives the most accurate of the estimated tested, significantly more so than the variance of periodicity.

The combined measure has a performance slightly worse than the best performing measure. Alternate ways of combining measures requires more investigation.

4.2 Multiple parameter tests

To test the feasibility of optimising multiple filter parameters, signal features were mapped as two separate dereverberation filter parameters are altered. Two results, for periodicity and LP residual kurtosis are shown in Fig. 5. All other reverberation measures discussed were tested in a similar manner, and display similar properties to those shown.

These representative plots show a single, smooth peak, close to the correct parameter, indicating optimisation methods should find the maximum value. The LPK contours are close to circular in shape, indicating low levels of dependence between the filter parameters. The periodicity contours show a diagonal shape indicating higher dependence. This means single parameter optimisation methods would need to iterate multiple times to find a higher dimensional peak, or a true multi-dimensional optimisation is required. A combined measure has a higher danger of being multi-modal, causing an optimisation method to find a local minimum.

4.3 Discussion

Ideally a full multi-dimensional parameter estimation would be performed. The estimated signal could then be compared to the original in both an estimation error and aural sense. However the amount of computation required would be prohibitive at this stage.

This test does provide a direct comparison between different features used for dereverberation, and a comparison between using full signals and those classified into appropriate segments. This tests whether speech classification is a useful tool for dereverberation and secondly which features (or combinations thereof) can be used to estimate the dereverberation filter.

The method the measures are combined over multiple frames also needs to be considered. Currently the mean is taken. An improved method which weights frames according to the energy of the frame may be a better method. This would reduce the difference between harmonic measurement over all frames and just voiced frames, as voiced frames typically contain the most energy.

5. CONCLUSION

This paper considered the framework of measuring a feature of a speech signal to determine the level of reverberation, and using this

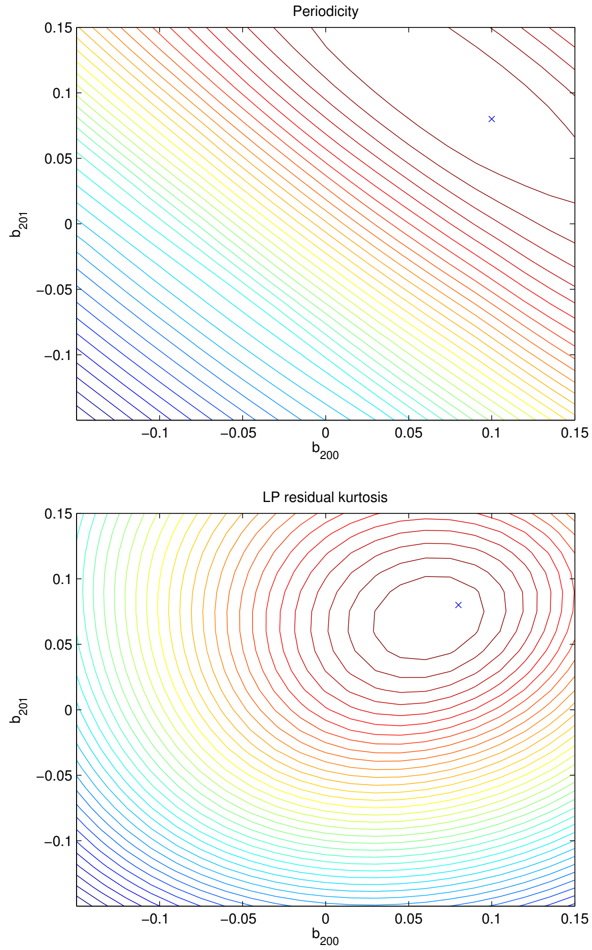

Figure 5: *Contour plot of periodicity (top) and LP residual kurtosis (bottom) of voiced segments for when two dereverberation filter parameters b_{200} and b_{201} are varied. Cross marks actual parameter values.*

measurement to estimate an inverse filter for dereverberation of the signal. One way to measure signal features is to classify segments of the speech into voiced, unvoiced and silence. This requires a speech classification method that works on reverberated speech, a subject on which little information could be found. Tests were performed, showing that classification performance degrades with increasing reverberation levels, but that some methods outperform others. Periodicity performs poorly at high reverberation levels, whereas ZCR performance drops at a slower rate. Outperforming both of these was a harmonic analysis based on a sinusoidal speech model.

Classification of silent periods allows the measurement of silence energy, which can be minimised to estimate a clean speech signal. Tests show this measure can give a good filter parameter estimate. Another new signal feature tested was signal entropy, which also gave a good parameter estimate. Entropy is very simple to calculate, with no classification step required.

Features which measure the harmonicity of voiced speech, including LP residual kurtosis, periodicity and harmonic energy, were also tested. The mean periodicity of voiced segments gave the most accurate estimate of measures tested, but required an accurate classification to work well. All these measures performed significantly worse when used over the entire signal rather than just voiced segments. This showed that a classification step can improve the performance of dereverberation methods which enhance these features.

REFERENCES

[1] P. J. Castellano, S. Sradharan, and D. Cole, “Speaker recognition in reverberant enclosures,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 117–120.

[2] P. S. Spencer and P. J. W. Rayner, “Separation of stationary and timevarying systems and its application to the restoration of gramophone recordings,” in *Proc. IEEE Int. Symp. Circuits and Systems*, May 1989, vol. 1, pp. 292–295.

[3] J. R. Hopgood and P. J. W. Rayner, “Blind single channel deconvolution using nonstationary signal processing,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 476–488, September 2003.

[4] T. Nakatani, K. Kinoshita, and M. Miyoshi, “Harmonicity-based blind dereverberation for single-channel speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 80–95, January 2007.

[5] B. W. Gillespie, H. S. Malvar, and D. A. A. Florêncio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 6, pp. 3701–3704.

[6] T. Nakatani, B. Juang, T. Hikichi, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, “Study on speech dereverberation with autocorrelation codebook,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. 193–196.

[7] T. Nakatani, B. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, “Importance of energy and spectral features in gaussian source model for speech dereverberation,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2007, pp. 299–302.

[8] T. F. Quatieri, *Discrete-time Speech Signal Processing. Principles and Practice*, Prentice Hall, 2002.

[9] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[10] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints for isolated utterances,” Tech. Rep. Vol. 54, No. 2, pp. 297–315, Bell System Technical Journal, 1975.

[11] R. Tucker, “Voice activity detection using a periodicity measure,” *IEE Proceedings*, vol. 139, no. 4, pp. 377–380, 1992.

[12] David H. Friedman, “Pseudo-maximum-likelihood speech pitch extraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 213–221, 1977.

[13] Robert J. McAulay and Thomas F. Quatieri, “Pitch estimation and voicing detection based on a sinusoidal speech model,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 1990, vol. 1, pp. 249–252.

[14] D. B. Paul, “The spectral envelope estimation vocoder,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1981, vol. 29, pp. 786–794.

[15] E. A. P. Habets, “Multi-channel speech dereverberation based on a statistical model of late reverberation,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 2003, vol. 4, pp. 173–176.

[16] J. R. Hopgood and C. Evers, “Block-based TVAR models for single-channel blind dereverberation of speech from a moving speaker,” in *IEEE Workshop on Statistical Signal Processing*, 2007, pp. 274–278.

[17] J. R. Fienup and J. J. Miller, “Aberration correction by maximizing generalized sharpness metrics,” *Journal of the Optical Society of America A*, vol. 20, no. 4, pp. 609–620, April 2003.