

Sparseness and speech perception in noise

Guoping Li, Mark E. Lutman

Institute of Sound and Vibration
University of Southampton, Southampton, Uk

lgp@isvr.soton.ac.uk

Abstract

Can we model speech recognition in noise by exploring higher order statistics of the combined signal? How will changes in these statistics affect speech perception in noise? This study addresses these questions in two experiments. One investigated the relationship between an established "glimpsing" model and the fourth order statistic, kurtosis. The glimpsing model [1] proposes that listeners can explore the local speech-to-noise ratio (SNR) in short time segments (glimpses) and focus on areas where SNR is high. Results showed that there is a very high correlation between percentages of glimpsing area and kurtosis ($r = 0.99; p < 0.01$), suggesting that kurtosis can serve as a simpler index for measuring glimpsing. The experiment also examined the association between kurtosis and recognition of nonsense words (vowel-consonant-vowel, VCV) in babble modulated noise, also showing very high correlation ($r = 0.97; p < 0.01$). Another separate study focused on the relationship of sparseness to speech recognition score for VCV words in natural babble noise made of 100 people talking simultaneously [2]. Results show that there is also high correlation between kurtosis and speech recognition score with this noise. Logistic regression analysis to obtain the kurtosis for 50% correct showed this was achieved at a kurtosis of approximately 1.0.

Index Terms: sparseness, speech perception, kurtosis

1. Introduction

It is known that the auditory system has many extraordinary abilities. One is the "cocktail party effect", where it can resolve many talkers speaking at the same time. The other is the feature of auditory filtering, which can be localized in both the time and frequency domains. Many theories and models [3] have been developed trying to account these phenomena. Few of them can explain these capabilities fully. Signal processing methods mimicking them by separating speech from noise have included information theory based models, such as independent component analysis (ICA). This sheds some light on these two important properties of the auditory system.

ICA explores higher order statistic of signals and can almost perfectly solve the instantaneous mixing problem, where two signals are mixed instantaneously (not convolved). This normally requires as many microphones as sources. Although this is still far from solving the practical cocktail party effect, it provides some clues for understanding how the auditory system may separate speech signals in this situation. Cooke [1] argues that listeners can actually get a good understanding of speech in noise by taking advantage time segments where the SNR is higher than the global SNR: the short-term segments are referred to as glimpses.

The idea of glimpsing is based on the known sparseness and

redundancy of speech, either in noise or against competing speakers. Speech signals are a highly modulated with many silences due to physical constraints on the speech production system. There tend to be many non-overlapped areas for speech mixtures when examined in the time-frequency domain. One assumption is that as the signal becomes more sparse, areas of clean speech increase and hence more opportunities for glimpsing occur. Listeners would then be able to reconstruct an accurate estimate of the whole speech signal through these glimpses and the redundancies in the speech. Cooke compared perceptual experiments of word (VCV) recognition in babble modulated noise [1] and the glimpse areas of these tokens. Results showed high correlation ($r = 0.955$) between the glimpse area of the VCV words and speech recognition scores of normal hearing subjects. A close fit between this glimpsing model and the perceptual evaluation results was observed. It proved that the area of glimpsing has a strong effect on the speech perception score, indicating that the subjects can take advantage of these areas for speech perception. It also suggests that the sparseness of speech in noise is important for speech perception in adverse environments.

There is also physiological evidence suggesting that sparseness is a key principle for neurons to encode environmental images and sounds. Everyday we receive large quantities of information, and our sensory system must have evolved efficient coding strategies to maximize the information conveyed to the brain without taking too many neural resources. Field [4] has shown that receptive field properties of simple cells in primary visual cortex produce a sparse representation. When this sparse representation is used as a constraint to encode images, a set of localized and oriented filters could be derived [5]. If applied to sound signals, a set of time and frequency localized filters can be derived [6, 7, 8]. These studies confirm sparse coding principles [9, 10] and the importance of statistics in neuroscience.

Based on the findings above it is reasonable to assume that, for speech perception in noisy environments, signals with more sparse structure will be easier to understand since it best fits the physiological encoding principle and should have greater glimpsing area. It follows that it is desirable to quantify sparseness in some way. Sparse speech signals necessarily have signal distributions that have more extreme peaks than Gaussian signals, due to the intermittency of production. A standard method to quantify sparseness is to use kurtosis, the 4th moment of the signal.

$$k = \frac{1}{n} \sum_{i=1}^n \frac{(r_i - \bar{\mu})^4}{\sigma^4} - 3 \quad (1)$$

where r is the amplitude of signal, $\bar{\mu}$ is the mean and σ is the standard deviation. For a Gaussian (non-sparse) distribution $k =$

0, whereas for non-Gaussian signals the kurtosis may be super-Gaussian ($k > 0$) or sub-Gaussian ($k < 0$).

Note that kurtosis is a simple property of the signal and may be computed much more easily than complex measures, such as glimpsing. This would be an advantage in real-time signal processing applications. The aim of the present study therefore was to assess the usefulness of kurtosis as a measure of sparseness, by examining its ability to predict speech recognition in noise.

2. sparseness optimization methods

Sparseness representation of data, means that the components of the representation are only rarely significantly active. Such representation is closely related to redundancy reduction and independent component analysis (ICA). Redundancy reduction can normally be achieved by Principal Component Analysis (PCA), which discards low-energy subspaces from highly dimensional data. ICA explores transform which can make the output as independent as possible. In this paper, the Projection Pursuit Algorithm was used to get a sparser signal out of the mixture of signals. Projection pursuit refers to the notion that the algorithm extracts the source signals one by one. Projection refers to a direction, which is orthogonal to all of the transformed sources signals except the one to be extracted (\hat{S}). Thus the inner product of mixtures of the signals and the unmixing matrix will produce one of the source signals.

2.1. The mixture

Sounds were mixed with babble noise by head related transfer function [11]. The noise was from 0 degrees and speech was from 90 degree. (Zero degrees means that the noise is in front, and 90 degrees means that the signal is on the right of the ear.)

2.2. Project Pursuit Gradient Ascent

This algorithms was described in detail by Stone [12]. It was introduced to separate two sounds by exploring sparseness of the signals, expressed by kurtosis. The unmixing matrix only updates when the kurtosis of the extracted signal is increasing; that is, only when the output is getting more sparse. The preprocessing by PCA (Principal component analysis) is also important; it makes the mean zero and normalizes the variance of mixture to 1. This is convenient for the calculation of kurtosis according to equation 1. The new equation for kurtosis will be:

$$k = E[(W^T X)^4] - 3 \quad (2)$$

The gradient of kurtosis for an extracted signal $Y = XW^T$ will be:

$$k(w)' = \alpha E[X(W^T X)^3] \quad (3)$$

where X is the mixture of signals, W is the unmixing matrix, Y is the extracted signal. α is a constant, and set it to unity 1 for convenience.

The unmixing matrix is updated by the gradient and old matrix:

$$W_{new} = W_{old} + k(w)' \quad (4)$$

And the unmixing matrix is normalized before the next iteration:

$$W_{new} = W_{new} / |W_{new}| \quad (5)$$

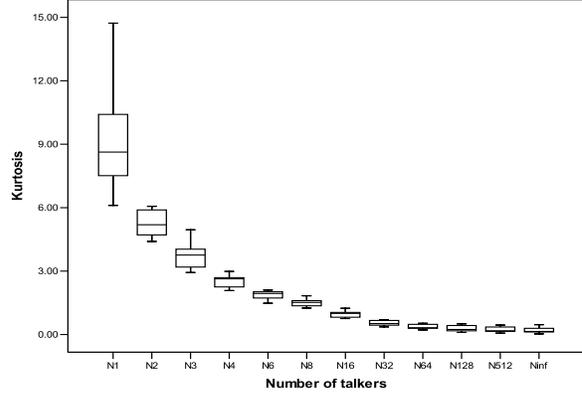


Figure 1: Kurtosis as a function of the number of talkers in babble modulated noise. 95% confidence interval is represented by the box. The line in the box represents the median. Maximum and minimum are indicated by the error bars.

According to equation 4, the kurtosis of the extracted signal for each iteration will be higher than that of signals extracted in the last iteration. A folder was created to save all the extracted signals and select some with the required kurtosis as test signals for experiment 2. The specific values of kurtosis were 0.32, 0.53, 1.08, 2.58, 4.

3. Experiments

Two experiments were performed to investigate the sparseness feature of speech and its relation to speech perception. The first experiment compared data from a published study using VCV words [1], in terms of glimpsing area, with the corresponding measure of kurtosis. The former measure has been shown to predict speech recognition scores accurately. The second experiment involved fresh measurements to investigate recognition of VCV words in babble noise, sorted with respect to the value of kurtosis in the range of 0 to 4.

3.1. Glimpsing and kurtosis

3.1.1. Speech materials

VCV words in babble modulated noise were used. These were the actual materials used by Cooke [1] and Simpson[13]. These included sixteen consonants (b, d, g, p, t, k, m, n, l, r, f, v, s, z, sh, tch) in the context of vowel /a/. The total test set includes 160 items from five male talkers and two examples of each talker were used. The noise signal was created by multiplying speech shaped noise with the long-term magnitude spectrum of the TIMIT corpus for various N. Twelve babble noise conditions were employed, $N = 1, 2, 3, 4, 6, 8, 16, 32, 64, 128, 512, \infty$. The noisy tokens were the sum of speech-shaped noise and 12 noise conditions at a global SNR of -6 dB.

3.1.2. Quantitative analysis of kurtosis

The kurtosis of each babble noise token as a function of N was calculated. Kurtosis was calculated based on equation (1) in the time domain of the each token. Fig.1 shows that kurtosis of the signal decreases continuously with the increase of babble talkers

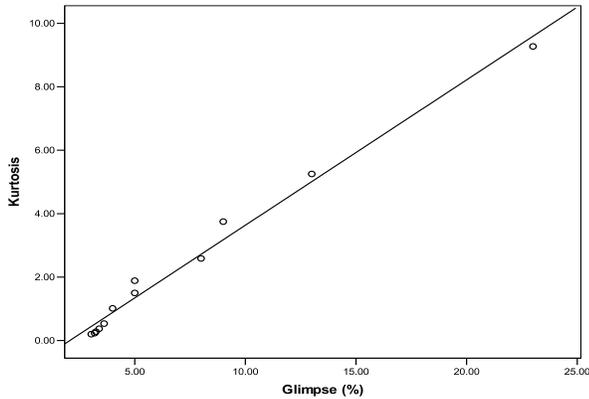


Figure 2: Correlation between kurtosis and glimpse area where the local SNR exceeds 3 dB. The results are averaged across all tokens and subjects. The correlation between kurtosis and glimpse area is 0.993 ($p < 0.01$). A signal with greater glimpse area is more sparse, with higher kurtosis.

N. It can be seen that spread of kurtosis is greater when there is only one talker. Student's t-test shows that there is no significance difference for $N > 16$. This is also true for the glimpse areas as observed in [1]. With the increase of N, the distribution approaches the Gaussian distribution, as indicated by the decrease of kurtosis.

3.1.3. Relationship between glimpsing and kurtosis

Fig. 2 plots the mean glimpse percentage in each noise condition against the corresponding kurtosis. The correlation is very high ($r = 0.993$). Cooke [1] proposed that glimpse area is a good predictor for the speech perception in babble modulated noise. He also found that the correlation between glimpse area and speech recognition score is high ($r = 0.955$). Our analysis using Cooke's recognition scores and our measure of kurtosis also showed a high correlation ($r = 0.97$; $p < 0.01$). These results suggest that kurtosis is a good predictor for glimpse area of speech tokens, and further a larger glimpse area indicates that a signal is more sparse.

3.2. Investigation of speech recognition in babble noise

In order to further investigate the relationship between kurtosis and speech understanding, new speech signals were produced in five groups with different kurtosis. An experiment was then performed with normal hearing subjects in a sound proof booth to obtain speech recognition scores as a function of kurtosis. Fig. 3 and Fig. 4 show examples of the sound /asa/, with different kurtosis. The kurtosis of the signals from S1 to S5 increases as a result of sparseness optimization.

3.2.1. Generation of speech and noise material

The set of 13 consonants b, d, f, g, j, l, m, n, p, s, t, v, z were used in the vowel context of /a/. These were available in the Matlab Nucleus toolbox, with single male talker. The babble speech noise was obtained from [2], which is the sound of 100 people talking in a canteen, with radius approximately two meters.

The speech and noise were mixed using a head related transfer function [11] with noise from 0 degrees and speech from 90 degrees. In this original mixture the speech was unrecognizable.

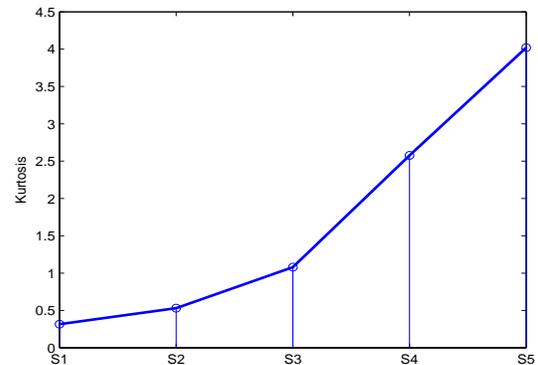


Figure 3: Example of kurtosis optimization for /asa/ sound in noise. The kurtosis of signal is increasing with different steps from S1 till S5. The waveform of each step is shown in Fig. 4

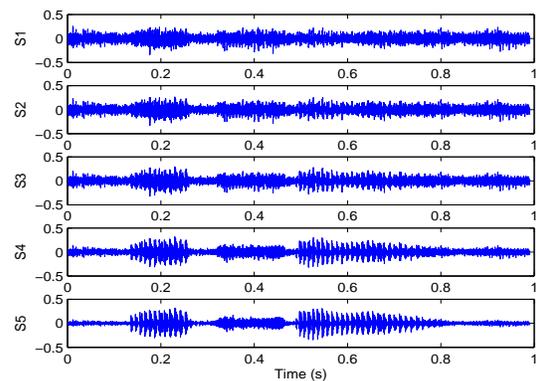


Figure 4: Examples of /asa/ sound waveform as a result of sparse optimization. From top to bottom, the kurtosis of the signal /asa/ are increasing (0.32, 0.53, 1.08, 2.58, 4), as shown in Fig. 3

To generate the signals required for the experiment, the kurtosis of signals was progressively increased using an iterative algorithm as shown in Fig. 3, by multiplying a matrix which updates according to an optimization principle. It updates only when the kurtosis of output increases. A series of signals were saved in a folder following each iteration and thus we produce a set of signals with increasing kurtosis. This is also called a projection pursuit algorithm [12]. Five groups of signals with different kurtosis were selected from the saved folders: $k = 0.38, 0.5, 1, 2.5, 4$.

3.2.2. Subjects and procedure

Seven listeners (3 male, 4 female) participated in the experiment. All had normal hearing. All listeners passed a pre-test on VCV words recognition in quiet at a recognition rate of at least 97%. The experiment took place in a sound isolated booth. Stimuli were presented monaurally through a TDH-39 earphone. Each subject completed five conditions with different kurtosis values. The initial 107 presentation for practice were not scored. Order of conditions and tokens were both randomized.

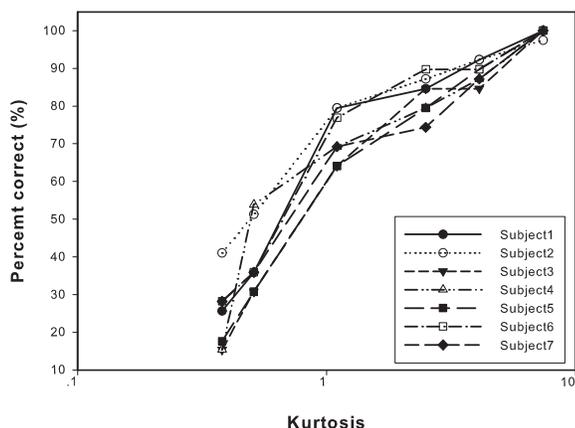


Figure 5: Consonant recognition in natural babble noise as a function of kurtosis, including clean speech ($k = 7.42$). The distribution is wider when kurtosis is smaller. The x-axis, kurtosis, is on log scale.

3.2.3. Results

Fig.5 plots the speech recognition score as a function of kurtosis, including clean speech. Speech recognition score and kurtosis have a high correlation ($r = 0.8, p < 0.01$). This shows directly that kurtosis is a good predictor for speech recognition in natural babble noise. It also shows that the recognition score increases very rapidly with the increase of kurtosis when the kurtosis value is low. The increase in score gets slower when the kurtosis reaches 2.5, which indicate that there is a ceiling effect. The kurtosis of the clean speech on average was 7.42. One-way analysis of variance (ANOVA) with post hoc comparisons of paired differences showed that the speech recognition scores for different kurtosis values were significantly different except for the pair $k = 4, 2.5$.

In order to further investigate the relation between the speech perception score and kurtosis, a logistic regression curve was calculated, as shown in Fig.6. The regression equation is:

$$P = \frac{1}{1 + e^{-(0.80 + 0.81x)}} \quad (6)$$

The regression curve emphasizes that the speech recognition score can be effectively predicted by the value of kurtosis, higher kurtosis predicting higher score. To get a 50% correct score, the kurtosis of such a signal should be no less than approximately 1.0.

3.3. Discussion

A signal with sparse representation can be considered to be biologically efficient. Sparseness is a key factor for the natural environment [5]. Our experiments have shown that kurtosis, a standard way of measuring sparseness, is a good predictor for speech perception in babble noise. These experimental results indicate that the sparseness feature of speech can indeed be exploited by the auditory system. Although this has been theoretically shown by many studies [7, 6, 14, 15, 8], our experiments show that this could be achieved by simply using the 4th order signal statistic, kurtosis.

Our study is limited to CVC words and could be usefully extended to other speech materials, such as sentences. According to

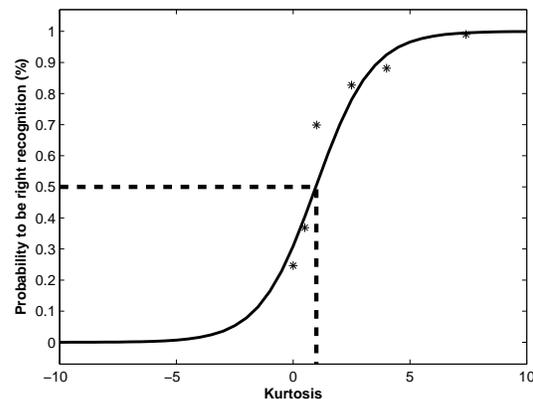


Figure 6: Logistic regression to show the relationship between speech perception and kurtosis. The observed recognition score from the second experiment is plotted as 'asterisks'. The point of 50% correct occurs when kurtosis equals 0.99.

the central limit theory, a mixture of signals is always more gaussian than each individual signal. So the kurtosis of the mixture is usually smaller than that of individual sources. A sparser representation of the mixture signal with higher kurtosis, would be closer to one of individual signals. And it would be easier for the speech recognition in noise, even for different words other than CVC. For sentences, there is the added complication of redundancy due to syntax and meaning. Further psychoacoustic experiments should be done to examine sentence recognition.

There are some shortcomings the way we have used kurtosis for measuring sparseness. First, it is very sensitive to the outliers. Considering equation (1), outliers may create significant changes for kurtosis. Second, kurtosis was calculated in the time domain only and focused on each CVC item as a whole. In the time-frequency domain, sparseness can be more easily understood as an indication of glimpsing areas where the SNR is high. However, an advantage of using kurtosis is that the calculation is direct and simple. A modified calculation could be used where the signal is divided into different time intervals and calculation of kurtosis could be based on a number of spectral regions. This might give a better predictor by considering the time-frequency domain characteristics of competing signals.

The kurtosis measure used in the present study can be calculated very simply and is amenable to simple updating using a sliding window. This makes it an attractive option for real-time applications than require an estimate of sparseness or prediction of speech recognition score. For example, kurtosis maximization can form the basis of algorithms for enhancement of speech in noise, such as used in modern digital signal-processing hearing aids. Repeated calculation of kurtosis would be computationally more efficient than more complex methods such as glimpsing.

3.4. Conclusion

Recognition of speech in babble noise is facilitated by sparseness in overlapping of the competing signals and by redundancy in speech. This sparseness can be represented simply by calculating kurtosis, with greater kurtosis corresponding to greater sparseness. This kurtosis measure is an effective predictor of speech recogni-

tion in babble noise, at least for vowel-consonant-vowel nonsense words. It compares well with other more complex methods, such as the glimpsing method. The simplicity of measuring kurtosis suggests it may find application in real-time signal-processing algorithms for the enhancement of speech in noise.

4. Acknowledgements

Thanks for Martin Cooke's generous help on providing data and discussion. Also thanks TNO Human Factors, Soesterberg, the Netherlands for providing the babble noise. This work is supported by a Rayleigh scholarship.

5. References

- [1] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, pp. 1562–1573, 2006.
- [2] Babble noise, "<http://spib.rice.edu/spib/data/signals/noise>," .
- [3] H. Purwins, B. Blankertz, and K. Obermayer, "Computing auditory perception," *Organized Sound*, vol. 5, pp. 159–171, 2000.
- [4] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells.," *J Opt Soc Am A*, vol. 4, pp. 2379–2394, 1987.
- [5] A. J. Bell and T. J. Sejnowski, "The independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, pp. 3327–3338, 1997.
- [6] M. S. Lewicki, "Learning optimal codes for natural images and sounds," *Proceedings of the SPIE - The International Society for Optical Engineering*, vol. 4119, pp. 185, 2000.
- [7] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, 2002.
- [8] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, 2006.
- [9] H. Barlow, "The knowledge used in vision and where it comes from," *Philos Trans R Soc Lond B Biol Sci*, vol. 352, pp. 1141–7, 1997.
- [10] H. Barlow, "Redundancy reduction revisited," *Network*, vol. 12, pp. 241–53, 2001.
- [11] B. Gardner and K. Martin, "Hrtf measurements of a kemar dummy-head microphone," 1994.
- [12] J. V. Stone, *Independent Component Analysis: A Tutorial Introduction*, Bradford Book, London, 1993.
- [13] S. A. Simpson and M. Cooke, "Consonant identification in n-talker babble is a nonmonotonic function of n," *The Journal of the Acoustical Society of America*, vol. 118, pp. 2775–2778, 2005.
- [14] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs.," *Curr Opin Neurobiol*, vol. 14, pp. 481–487, 2004.
- [15] B. A. Olshausen and K. N. O'Connor, "A new window on sound," *Nat Neurosci*, vol. 5, pp. 292–294, 2002.