FAVOURABLE AND UNFAVOURABLE SHORT DURATION SEGMENTS OF SPEECH IN NOISE

Daniel Woo

School of Electrical Engineering

The University of New South Wales

ABSTRACT

Human perceptual experiments are described that present listeners with segmented stop consonant speech stimuli in noise. The selection of short duration speech segments is based on a local measure of the signal-to-noise ratio calculated over 1ms windows. The aim is to create stimuli with known fluctuations occurring between a speech and noise sample to assess whether the presence of short duration "gaps" in the noise produce favourable and unfavourable signal regions that influence identification. Perceptual results are reported that suggest human listeners make better use of signals that comprise only of positive, local signal-to-noise ratio segments. Such regions are assumed to be more favourable for stimuli identification. Presentation of stimuli containing only negative signal-to-noise ratio regions does not appear to contribute as much. A model that is based on the accumulation of short duration spectral segments is presented that produces a similar set of identification functions for the same test stimuli.

1. INTRODUCTION

When speech and noise are additively combined, a single global measure of the signal-to-noise ratio (SNR) is generally used to define the overall energy relationship between the two signals. For example, vowels contain more energy compared with unvoiced stop consonants and a global SNR does not reflect how low energy speech segments are more adversely affected compared with high energy voiced sounds. The question under consideration is how short duration, time-varying fluctuations in the SNR affect identification performance. Only stop consonants are considered in this study.

2. BACKGROUND

Miller and Licklider (3) considered the effect of fluctuating background noise on the perception of continuous speech material by measuring human identification performance in the presence of regularly and randomly interrupted noise stimuli. Various SNR (+9 to -18dB) and interruption rates were considered from 0.1Hz through to 10kHz. Identification performance with high interruption rates was considered similar to continuous noise and the relationship between the interruption rate and the syllable duration was believed to be a contributing factor.

Studies by Howard-Jones et al. (5, 4) have considered the effect of interruptions in both time and frequency using noise with a time-frequency distribution that looks like a checkerboard. They have reported that fluctuations in both time and frequency do produce statistically different identification results. Their proposal suggests that humans can make use of time-frequency "gaps" in the noise. Howard-Jones experiments have only considered interaction using regular time interruptions of 10Hz.

The concept of time-varying fluctuations due to the phase of a noise signal has been suggested in (7) and investigated in recent perceptual studies by Summers and Leek (8). These studies suggest that over short periods, the masking effectiveness of noise varies. This supports the notion that short gaps in the noise contribute to improved performance. Summers et al. (8) have shown using an auditory model that variations in the phase of signals produce different basilar membrane responses.

3. SIGNAL SEGMENTATION

In this series of experiments (10) six initial position stop consonants /b,d,g,k,p,t/ are considered in the presence of non-stationary office noise samples (6). Only the first 30ms from the burst release of the stop consonants is used in the speech material (Figure 1).

Three types of signals are generated for a given global SNR, "speech plus noise" (**S+N**), "speech above plus noise" (**A+N**) and "speech below plus noise" (**B+N**). The first signal (**S+N**) is the typical case of combining the speech signal with noise at a specified SNR. The other two combined signals require measurement of the local RMS value for each signal (Figure 2) calculation of the local SNR (Figure 3a) using non-overlapping 1ms windows at the same gain scaling factor used to generate

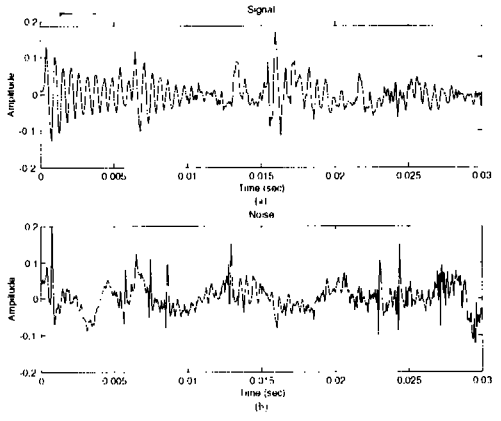

Figure 1: (a) Speech signal and (b) office noise signal.

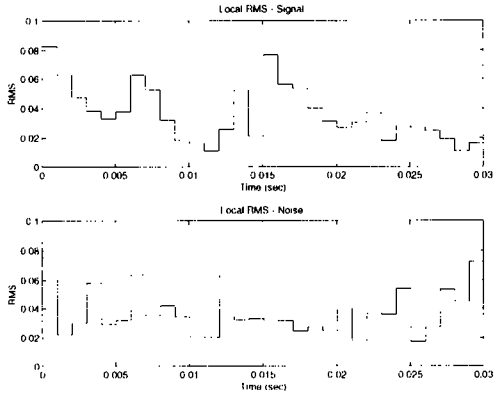

Figure 2: RMS value for the signal and noise measured over 1ms intervals.

S+N. A mask is created based on whether the local SNR is positive or strictly negative (Figure 3b). The mask is used to extract regions of the speech signal with positive SNR (Figure 4) which are subsequently combined with a continuous noise sample. The third signal (**B+N**) is generated in an analogous manner to the **A+N** signal except that the complement of the mask is used to select the speech regions. Unlike (3), the speech signal is segmented rather than the noise and much shorter durations are considered compared with (5).

4. PERCEPTUAL EXPERIMENTS

4.1. Method

A stop consonant speech database consisting of two male and two female speakers was prepared by excising the first 30ms of initial position speech tokens /b,d,g,k,p,t/ from continuous

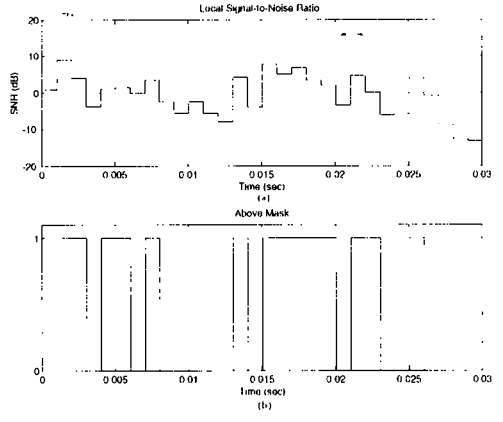

Figure 3: (a) Local signal-to-noise ratio. (b) Above mask.

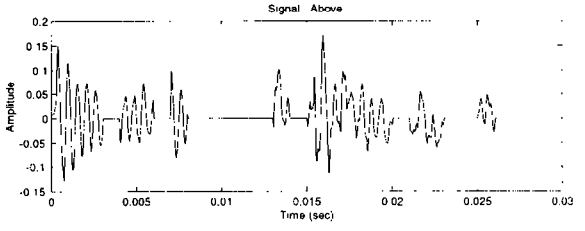

Figure 4: Above signal prior to being mixed with noise.

speech sentences in the same vowel and word contexts. All speakers had Australian English accents.

Clean speech tokens were combined with three office noise samples using three mixing conditions (**S+N**, **A+N** and **B+N**) at four SNRs +2, 0, -2 and -4dB. Two blocks each containing the complete set of stimuli were presented using different randomisation patterns.

The speech material was presented to a group of 20 listeners consisting of approximately equal numbers of male and female subjects with normal hearing. Experiments were conducted in a sound treated room with audio presented through a Madsen audiometer distributed to TDH-39 headphones. Presentation levels were 70dB SPL. Subjects were presented with an utterance and then entered their response using a touch screen button. Voiced and unvoiced experiments were conducted separately and listeners only selected one from three available choices.

4.2. Results

Identification results for both voiced and unvoiced experiments are shown in Figure 5. The **S+N** case produces the best overall

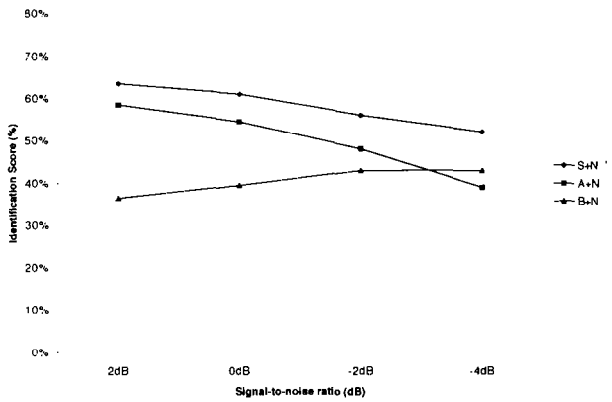

Figure 5: Identification score for identifying the stop consonants (both voiced and unvoiced) for different signal mixing conditions and signal-to-noise ratios.

performance and represents the case for both continuous noise and continuous speech. In keeping with expectations, as the global SNR is decreased, identification performance also decreases. For the mixing condition **A+N**, where only positive SNR regions are included, the identification score lies slightly below the **S+N** case at +2 and 0dB global SNR. The difference between **S+N** and **A+N** becomes greater at negative global SNRs. This arises from the relationship between the global SNR and the number of speech regions included in the signal. As the global SNR becomes more negative, the total number of local positive SNR regions decreases. This can be visualised by considering Figure 3a. If the dotted line is considered as the threshold, then at negative SNRs the local SNR function (solid line) shifts downward and the number of regions appearing above the line decreases. At the same time, the number of "below" regions increases. This is observed in the **B+N** case (Figure 5) where decreasing global SNR yields increasing performance. The **A+N** performance falls below that of **B+N** at -4dB because the total duration of speech in the **A+N** signal is quite small.

An interesting point to note is that at 0dB SNR where approximately equal durations of speech are present, the identification of the **A+N** condition is significantly greater than **B+N**, suggesting that although partial segments of the speech signal are presented, listeners perform better when presented with only those regions of positive local SNR.

5. A TIME-FREQUENCY MODEL

A pattern classification model is proposed (10, 11) that is based on the observed behaviour of human listeners when presented with successively longer stimuli durations using 1-5ms increments. As the total stimuli duration increases human performance also increases (9,1). Presentation of only the short duration increments (<10ms) in isolation yields chance performance (1). One interpretation of these findings is that speech perception is facilitated by the amalgamation of multiple short duration spectral representations.

The perceptual results of this study indicate that humans can make use of non-contiguous regions of positive local SNR more effectively than signal containing negative regions. The speech stimuli are segmented and can contain speech events with minimum duration of 1ms or longer. This presents a challenge to conventional speech processing techniques since typical analysis window durations of 10-20ms are unable to reveal the short duration events present in the signal. Fourier techniques using 1ms windows are avoided to circumvent the loss of frequency resolution.

A positive time-frequency representation (2) is used to represent the speech and noise signal and has been found to represent short duration speech events more accurately than Fourier and Wigner techniques avoiding the time-frequency trade-off and both artefact and negative components (10).

A spectral comparison is made against a set of stored clean templates for each 1ms spectral slice derived from the positive time-frequency representation. An histogram of the set of best-matching tokens is formed for each 1ms interval and integrated over the 30ms interval (Figure 6). Templates are derived from clean speech stimuli while test utterances are produced by combining the clean stimuli with noise.

5.1. Results

By applying the same speech stimuli used in the perceptual experiments, the identification results produced by the model retain similar characteristics observed for human performance (Figure 7). The identification score decreases for negative global SNR for **S+N** and **A+N**, the **A+N** score deviates more from **S+N** at negative SNRs indicating the decreased amount of speech. **B+N** rises as the SNR is decreased although not as much as the perceptual results and at 0dB SNR the **A+N** condition is better identified compared with **B+N** despite containing approximately the same durations of speech.

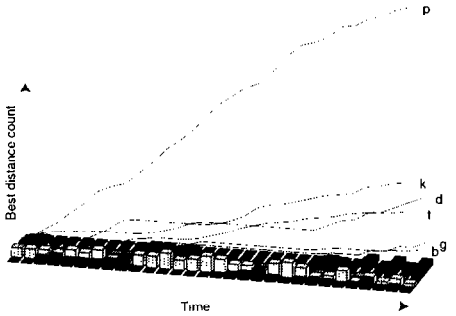

Figure 6: Histograms for each 1ms interval (3D bar graph) with the integrated result for each token. /p/ is the correct token in this example.

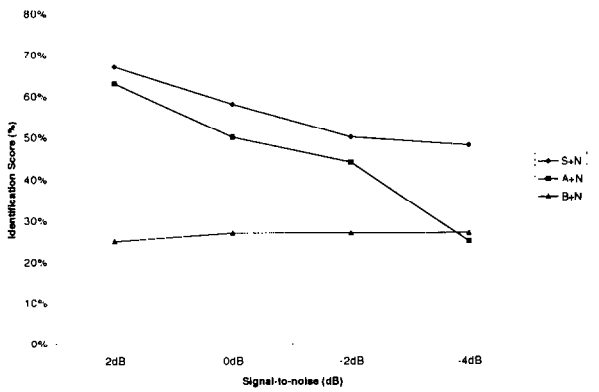

Figure 7: Identification performance produce by the model

6. SUMMARY

The time-varying properties of a speech and noise signal produce fluctuations in the local signal-to-noise ratio. Presentation of only the positive SNR regions appear to offer more favourable conditions for the identification of speech compared with regions of negative SNR. Using a model that integrates short duration (1ms) spectral information derived from a positive time-frequency distribution, identification performance has been produced that supports the observations found for human listeners. The model agrees with the hypothesis that short duration regions of positive SNR are more beneficial for identification.

7. ACKNOWLEDGEMENTS

I would like to thank Dr. Phillip Dermody and Dr. Chris Phillips for their support and guidance during the evolution of this work.

8. REFERENCES

1. Dermody, P. J., *Perceptual Processing of Stop Consonants as Initial Sounds in Spoken English*, Ph.D. Thesis, The University of New South Wales, January 1989.

2. Loughlin, P.J., Pitton, J.W., and Atlas, L.E., *Construction of Positive Time-Frequency Distributions*, IEEE Trans. Sig. Proc., 42(10), pp.2697-2795. October 1994.

3. Miller, G. A. and Licklider, J.C.R., *The Intelligibility of Interrupted Speech*. JASA, 22 (2), pp.167-173, March 1950.

4. Howard-Jones, P. A. and Rosen, S., *The Perception of Speech in Fluctuating Noise*. Acustica, 78, pp.258-272, 1993.

5. Howard-Jones, P. A. and Rosen, S., *Uncomodulated glimpsing in "checkerboard" noise*. JASA, 93 (5), pp.2915-2922, May 1993.

6. Raicevich, G., *Noise Characterisation of Office Environments for Speech Recogniser Operation*, SCRG Technical Report, National Acoustics Laboratory, Chatswood. Australia, June 1991.

7. Schroeder, M. R. and Mehrgardt, S., *Auditory Masking Phenomena in the Perception of Speech*, R. Carlson and B. Granstrom (Ed.), The Representation of Speech in the Peripheral Auditory System, Elsevier Biomedical Press, 1982.

8. Summers, V. and Leek, M.R., *Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners*, Hearing Research, 118, pp.139-150, 1998.

9. Tekieli, M. E. and Cullinan, W. L., *The Perception of Temporally Segmented Vowels and Consonant-Vowel Syllables*, Journal of Speech and Hearing Research, 22(1), pp.103-121, March 1979.

10. Woo, D. T., *Human Perception of Stop Consonants in Noise: A Time-Frequency Model*. Ph.D Thesis, The University of New South Wales, Australia. July 1998.

11. Woo, D. T., Dermody, P.J., and Phillips, C.J.E., *Analysis of Human Gating Performance Using Time-Frequency Analysis*, Proc. 1996 ANZIIS, Adelaide, Australia. pp.117-119. November 1996.