An information theoretic tool for investigating speech perception

Bryce Lobdell, Jont B. Allen

Department of Electrical and Computer Engineering
University of Illinois, Urbana, Illinois
lobdell@uiuc.edu, jontalle@uiuc.edu

Abstract

A method for investigating human speech perception that combines information about human perception of speech, auditory modeling, signal detection theory, and information theory is described. For that purpose, a model for detection of signals in the auditory nerve is developed and used to analyze speech sounds. Examples are given that combine detectability information with $P(\text{heard} = h|\text{spoken} = s)$ to access information about "cues" to some some speech sounds. These example agree qualitatively with previous results about perception of these sounds. Refinements and prospects for this approach are discussed in light of the examples.

Index Terms: speech perception, information theory, signal detection theory, auditory modeling, auditory nerve.

1. Introduction

Despite impressive improvements in machine speech recognition performance and ingenuity in the design of machine speech recognizers, machine speech recognition does not have humanlike robustness to noise [1], and other types of degraded listening conditions. Human speech recognition can tolerate spectral degradation [2], temporal distortion [3], noise and filtering [4], and even more invasive modification such as conversion to sine wave speech [5] with marginal to moderate loss of performance, whereas machine speech recognition performance suffers greatly under virtually any adverse conditions. The remarkable robustness of human speech perception inspires us to investigate human speech perception, particularly feature extraction [6], with the hope that more appropriate feature extraction could improve the performance of machine speech recognition.

The purpose of this paper is to describe a tool that can be used for investigating human speech perception. It has been a theme of speech perception research to search for "cues" to "units" in speech perception [7], but those discoveries have had limited application in machine speech recognition research. It is hoped that the tool described here can help us determine the structure of the brain's speech recognition hardware using speech perception experiments. However in this paper, the tool will be used to (re)discover "cues" to some consonant sounds as a way of testing the usefulness of this tool. If the tool appears to be a reasonable model, it could be further developed to give information not just about "cues" to "units" in speech perception but also about perceptually relevant neural representations of speech. Perhaps that would provide a more appropriate parameterization of speech for machine speech recognition than the linear prediction coefficients or the Mel frequency cepstra often used in machine speech recognition.

The articulation index, which was developed to predict the intelligibility of speech sounds [8], connects information theory and speech perception performance via a formula that provides the probability of correctly identifying nonsense speech sounds P_c in terms of the filtered speech spectrum and the interfering noise spectrum [9]. The formula for P_c resembles the Shannon channel capacity in that it involves a term $\log(1 + c^2\sigma_s/\sigma_n)$, which is the logarithm of the scaled number of just-noticeable-differences (JNDs), which itself has the interpretation of information.

In an experiment by Miller and Nicely [10] test subjects were asked to identify (from a closed set of responses) consonant-vowel sounds that had been filtered and mixed with noise. The listener responses revealed perceptual categories, which presumably have common (statistically variable) acoustic correlates that can be investigated with the help of detectability information. Allen [11] has also shown that there is a relationship between the articulation index and the confusion patterns of Miller and Nicely.

These insights encourage us to leverage closed set speech perception experiments involving distorted speech, along with auditory modeling, and a detection model to investigate speech perception, as shown in Fig. 1. The auditory and detectability model is a measure of input, and a confusion matrix formed from human listener responses is a measure of the output.

A closed and resonably small set of nonsense (maximum entropy) speech sounds is necessary for this type of informaton theoretic analysis. A speech perception experiment involving real words or natural speech does not allow the analysis of particular phonetic distinctions, and it involves syntactic, lexical, and symantic constraints that would confound our efforts to find the source of the phonetic distinctions. This type of experiment can provide information about the perception of continuous speech if it can be used to discover relevant neural representations of signals in the auditory system that are always used for speech perception.

Section 2 describes a model for determining detectability at the auditory nerve level of acoustic signals in noise. Section 3 describes an attempt to find "cues" to perceptual categories using the auditory model and detection model described in Section 2.

2. Auditory Model and Detection Model

This section describes a method for simulating signals in the auditory system and determining the detectability in distortion of those signals for a particular definition of detectability. The distortion of interest in this case is stationary noise as was used in [10], which has the purpose of masking "cues."

The detection model will tell us where, in time and frequency,

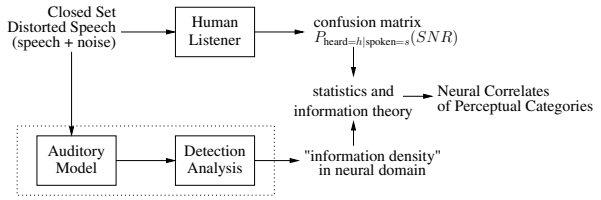

Figure 1: The proposed method for investigating human speech perception. The blocks enclosed in the dotted line are described in section 2.

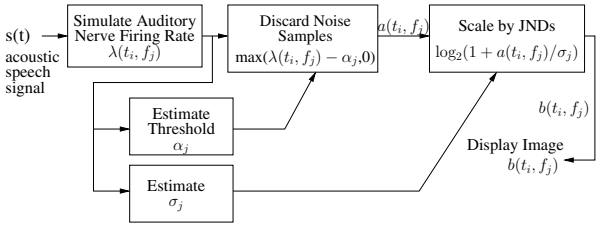

Figure 2: Block diagram of the system used to determine when the auditory nerve firing rate is detectable to the auditory system.

an acoustic signal will cause the auditory nerve firing rate to be detectable to the auditory system, and thus available for speech recognition. The result will be an image with one index corresponding to the best frequency of the auditory nerve, and the other to time.

The system is shown in Fig. 2. The first stage is to simulate the auditory nerve firing rate for for the acoustic signal. This can be done using virtually any auditory model (for example, [12], [13]). Initially we used a set of linear filters that provide correct tone detection thresholds. A set of frequencies f_j are chosen, which are the "best frequency" for the collection of auditory nerves we wish to simulate. The frequencies were chosen using the Greenwood map so that they would be spaced evenly along the basilar membrane.

Next, the probability distribution of the auditory nerve firing rate resulting from the noise is determined. The threshold α is determined based on that probability distribution for every frequency f_j simulated. The threshold α_j is chosen so that the maximum likelihood detector will detect a tone with the roughly the same probability as a human listener at a variety of signal-to-noise ratios (SNRs) near the detection threshold for humans. Figure 3 shows an example auditory nerve rate due to a speech sound (/ʊɪ/) mixed with noise, along with a cartoon of the noise and noise plus speech distributions.

The articulation index is the weighted sum across frequency of $\log(1+c^2\sigma_s/\sigma_n)$ [11]. The quantity $c^2\sigma_s/\sigma_n$ is interpreted as the number of JNDs due to noise, and $\log_2(1+c^2\sigma_s/\sigma_n)$ is interpreted as the number of bits per cycle that can be conveyed with that number of JNDs. This relationship between the JND, the amount of information available, and the probability correct provided by the articulation index inspires us to scale the auditory and detection model accordingly. The standard deviation of the noise in each auditory nerve channel σ_j is computed from the noise distribution. The samples from the auditory nerve firing rate $\lambda(t_i, f_j)$ below the corresponding threshold are set to zero, and the samples above the threshold are scaled according to $\log_2(1+a(t_i, f_j)/\sigma_j)$, as shown in Fig. 2.

The signal $b(t_i, f_j)$, which is the scaled version of the signal $a(t_i, f_j)$ is then displayed as an image. It bears a resemblance to the spectrogram, except that the time-frequency resolution is

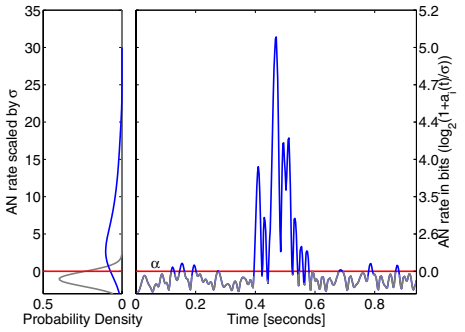

Figure 3: The right pane shows an example of the firing rate for an auditory nerve with best a best frequency of 1 kHz induced by a speech sound at a wideband SNR of zero decibels. The horizontal line at zero on the ordinate is the threshold α. The left pane shows the noise distribution and noise plus speech distribution of the auditory nerve firing rate in that channel.

similar to that of the auditory nerve, the spacing of the frequency channels is uniformly spaced on the human basilar membrane, and the regions (in time and frequency) below the "noise floor" are blank. The next section presents some of these images, referred to as the CCGRAM (channel capacity-gram), at several SNR, the corresponding spectrogram, and the human confusion associated with those images.

3. Usage and Comparison with a Spectrogram

Figure 4 shows human confusions of a particular recording of the consonant-vowel pair /ʊɑ/ (from /v/acuum and f/a/ther). The abscissa is the SNR for the sound presented. Each line shows the probability of this recording of /ʊɑ/ being perceived as other sound as a function of SNR. The data shown in Figs. 4, 5, and 7 were collected in an experiment which duplicates the Miller and Nicely experiment as closely as possible [14]. There were 16 consonant sounds paired with /ɑ/, and white interfering noise, and 16 response alternatives corresponding to the 16 consonants used.

Like most recordings of /ʊɑ/, this one has energy that precedes the transition into the vowel (the so-called voice bar) /ɑ/, which occurs between 0.1 and 0.2 s and between 0.5 and 2 kHz, shown (most clearly) in the left pane of Fig. 6. Perception of /ʊ/ is generally attributed to the presence of this voice bar, the place of articulation identified by the spectral shape of the region above the voice bar, and the formant trajectory [15]. This particular recording of /ʊɑ/ is always perceived as /ʊ/ down to 12 dB SNR, recognition as /ʊ/ decreases slowly between 12 dB SNR and -6 dB SNR. Below -6 dB SNR, recognition of /ʊ/ diminishes rapidly.

Figure 6 shows the CCGRAM in the upper panes and a narrow-band spectrogram in the lower panes, corresponding to SNRs of 12, -3, and -9 from left to right. The CCGRAM shows that the voice bar is clearly visible at 12 dB SNR, barely visible at -3 dB SNR, and invisible at -9 dB SNR after recognition of /ʊ/ has decreased rapidly. More comparison between the CCGRAM and the confusion pattern pattern for /ʊ/ and other sounds that are similar, such as /ƀ/ and /f/, provide better than circumstantial evidence for the perceptual importance of the voice bar for recognition of /ʊ/.

Next, the CCGRAM is used to describe properties of a speech sound that are responsible for a certain distinction. Some utter-

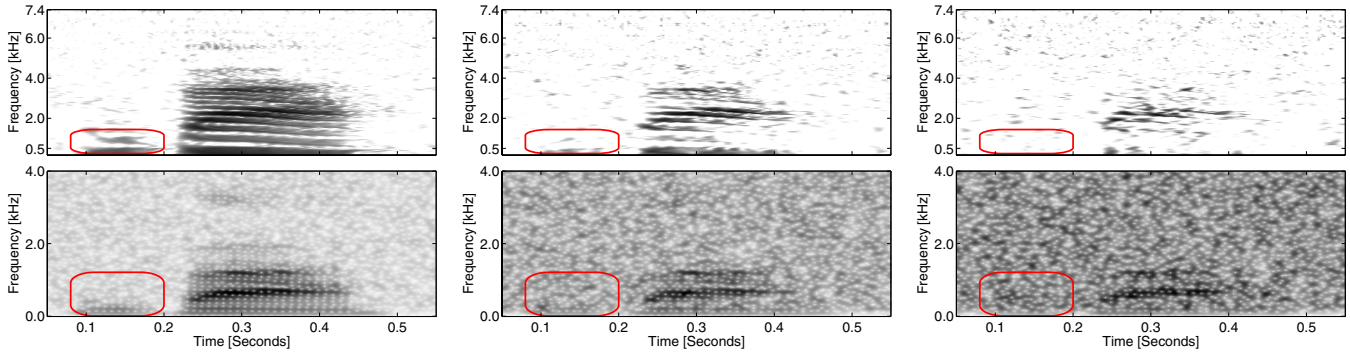

Figure 6: The CCGRAM (top) and narrow-band spectrogram (bottom) of a particular utterance of the speech sound /ʋ/ mixed with noise. The wideband RMS-based SNR is 12 dB for the left pane, -3 dB for the middle pane, and -9 dB for the right pane.

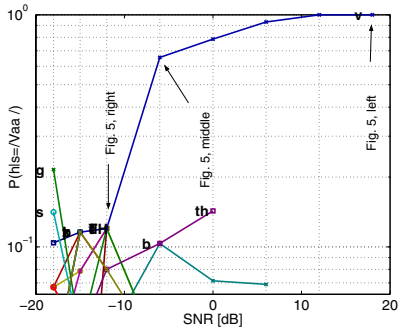

Figure 4: Confusion pattern for a particular utterance of the consonant-vowel pair /ʋɑ/. The abscissa is the SNR (in white noise) for the sound as presented to a human listener. The ordinate is the probability of recognition for each of 16 choices, each choice having a different line.

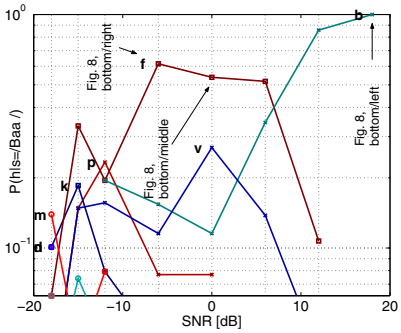

Figure 5: Confusion pattern for a particular utterance of the consonant-vowel pair /bɑ/. The abscissa is the SNR (in white noise) for the sound as presented to a human listener. The ordinate is the probability of recognition for each of 16 choices, each choice having a different line.

ances of /bɑ/ are confused with /ʋ/ when mixed with white noise, and others are confused with /f/. Figures 5 and 7 show the human confusions for two utterances of the sound /bɑ/ that exhibit this behavior. At high signal-to-noise ratios both sounds are perceived as /b/, but at low SNRs they are perceived as /ʋ/ and /f/, respectively. The left pane of Fig. 8 shows the CCGRAM of the recordings of /bɑ/ at a high SNR, which exhibit the familiar rising first and second formants, which are thought to cue /b/ when followed by /ɑ/. The middle pane of Fig. 8 shows the same sounds at 0 dB SNR, at

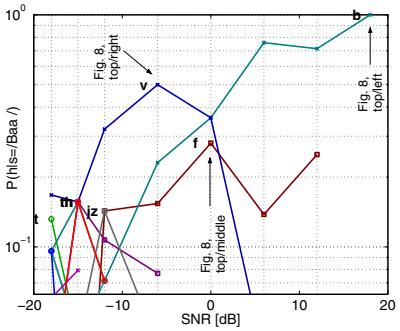

Figure 7: Confusion pattern for a particular utterance of the consonant-vowel pair /bɑ/. The abscissa is the SNR (in white noise) for the sound as presented to a human listener. The ordinate is the probability of recognition for each of 16 choices, each choice having a different line.

which point there is little information outside the darkened region. The right pane of Fig. 8 show the same sounds a -6 dB SNR, at which point there is no information available outside the darkened region. The conclusion is that information within that region is causing one sound to be perceived as /f/ and the other as /ʋ/. This agrees qualitatively the with result of [16], which asserts that the starting frequency of the first formant and other fine details about the formants cue "voicing" for stop consonants. In fact, inspection of the CCGRAMs of more utterances of /b/, /ʋ/, and /f/ agree qualitatively in that a lower starting frequency of the second formant makes the sound more likely to be perceived as /b/ or /ʋ/, rather than /f/.

4. Discussion and Conclusions

The first example involving Figs. 4 and 6 shows an example where the probability of recognition as /ʋ/ is related to the "information" in a particular region of the CCGRAM. This example is representative of the utterances of /ʋ/ available to us, and also in agreement with long standing observations about speech perception.

The second example, involving Figs. 5 and 8, shows how the CCGRAM can be used to specify the location (in time and frequency) of the "cue" that leads to a certain distinction.

These arguments, while not supported here by statistically sufficient evidence (which would be more confusion patterns and associated CCGRAMs), illustrate the type of approach that can be

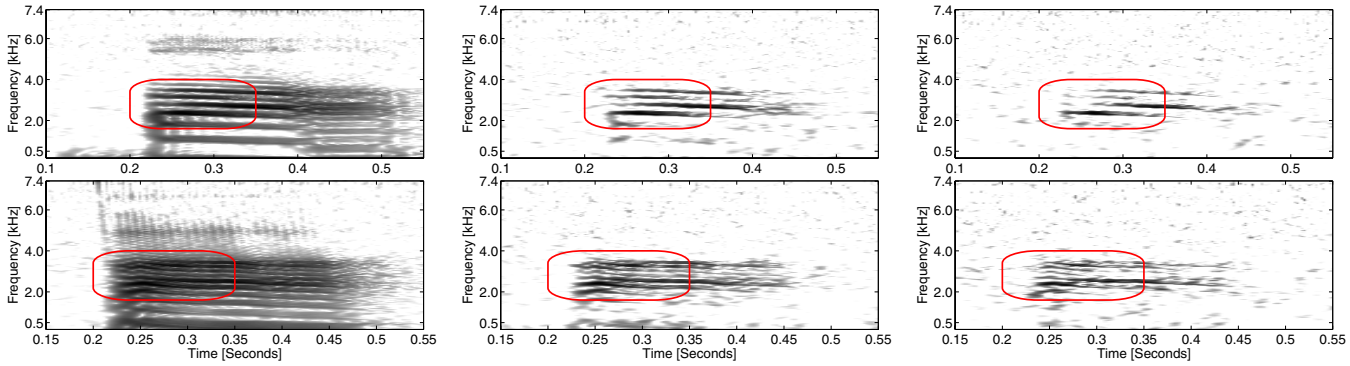

Figure 8: The CCGRAM of two recordings of the sound /bɑ/ spoken by two different people at a SNR of 24, 0, and -6 from left to right. The top pane shows the sound from Fig. 7, the bottom pane shows the sound from Fig. 5. In the leftmost pane, both sounds are perceived as /b/ with a probability of 100%. In the middle pane, both sounds are somewhat ambiguous between /b/, /v/, and /f/. In the rightmost pane, the sound corresponding to the top CCGRAM is perceived as /v/ with a probability of 50%, and the sound corresponding to the bottom CCGRAM is perceived as /f/ with a probability of 60%.

used to connect a measure of the output of human speech perception, the confusion matrix, with a measure of the input information, the CCGRAM. The spectrogram does not provide detectability information because it is not based on a physiologically plausible model of signals in the auditory system and because it does not explicitly deal with noise and its effect on detectability.

The usefulness of this approach for investigating speech perception could be extended by devising a way to automatically connect the detectability information with the human confusion information. Joint and conditional probability density functions between the listener responses and the detectability information from the CCGRAM could be estimated and used to compute the mutual information (or some other information theoretic measure) between perceptual categories and the acoustic information for categories of speech sounds.

We hope that methods can be devised that would provide information about fine phonetic detail, as well as provide statistics on the relevance of those details. It is also hoped that these methods can be used to access information about which signals in the auditory system are relevant to speech perception.

5. Acknowledgments

We would like to thank Stephen E. Levinson and Robert Wickesburg for their helpful input. And also the members of the HSR group for their related work.

6. References

[1] R.P. Lippman, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[2] R.V. Shannon, F-G. Zeng, V. Kamath, J. Wygonski, and M. Ekclid, "Speech recognition with parimarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, October 1995.

[3] Plomp R. Drullman R., Festen J., "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[4] Steinberg J.C. French N.R., "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90–119, 1947.

[5] R.E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, pp. 947–950, 1981.

[6] J.B. Allen, "How do humans process and recognize speech?," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[7] Sarah Hawkins, "Puzzles and patterns in 50 years of research on speech perception," in *From Sound to Sense*, June 2004.

[8] Harvey Fletcher and R.H. Galt, "Perception of speech and its relation to telephony," *J. Acoust. Soc. Am.*, vol. 22, pp. 89–151, Mar. 1950.

[9] J.B. Allen, *Articulation and Intelligibility*, Morgan and Claypool Publishers, 2005.

[10] Nicely P.E. Miller G.A., "An analysis of perceptual confusions among some english consonants," *J. Acoust. Soc. Am.*, vol. 27, no. 2, pp. 338–352, March 1955.

[11] J.B. Allen, "Consonant recognition and the articulation index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2212–2223, April 2005.

[12] J.B. Allen, "Nonlinear cochlear signal processing," in *Physiology of the Ear, Second Edition*, A.F. Jahn and J. Santos-Sacchi, Eds. Singular Thomson Learning, 401 West A Street, Suite 325 San Diego, CA 92101, 2000.

[13] X. Zhang, M.G. Heinz, I.C. Bruce, and L.H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 648–670, February 2001.

[14] Andrew Lovitt, Sandeep Phatak, and Jont Allen, "Interpreting consonant and vowel confusion functions using information-theoretic measures.," in *Of the twenty-ninth Midwinter Research Meeting of the Association for Research in Otolaryngology.*, 2006.

[15] Kopp H.G. Potter R.K., Kopp G.A., *Visible Speech*, Dover Publications, Inc., New York, New York, 1966.

[16] J. Jiang, M. Chen, and A. Alwan, "On the perception of voicing in syllable-initial plosives in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 2, pp. 1092–1105, Feb. 2006.