# Human Speech Perception and Feature Extraction

*Bryce E. Lobdell, Mark A. Hasegawa-Johnson, Jont B. Allen*

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

## Abstract

Speech perception experiments tell us a great deal about which factors affect human performance and behavior. In particular many experiments indicate that the signal-to-noise ratio spectrum is an important factor, indeed the signal-to-noise ratio spectrum is the basis of the Articulation Index, a standard measure of "speech channel capacity." In this paper we compare speech recognition performance for features based on the Articulation Index with two alternatives typically used in speech recognition. The experimental conditions vary the spectrum and level of noise distorting the speech in the training and test set. The perceptually inspired features generally perform better when there is a mismatch between the training and test noise spectrum and level, but worse when the test and training noises match.

## 1. Introduction

Speech perception by humans is often studied in the context of pattern recognition: Signals are presented to humans and the humans make some objective judgment about what they heard. In some cases the signal is parametrically modified synthetic speech, in others it is natural speech which has been distorted. A great deal of descriptive knowledge exists about human behavior in speech perception; we propose to leverage that knowledge to improve automatic speech recognition.

One focus of speech perception research has been the design of speech communication equipment or auditoria for speech intelligibility. The most exhaustive of such studies resulted in the *Articulation Index* (AI), which is a method for predicting the probability of correct phone identification by humans from purely physical parameters of the speech communication channel. Some of the parameters are masking (by noise or periodic sounds), filtering, and sound level. The AI was formulated by modifying those (and other) parameters of a speech communications channel, empirically measuring the resulting error probability (using a human "reader" and "listener"), and modeling the relationship between physical parameters and score. The result of this effort, the Articulation Index, is published in [6, 5, 8, 9, 11, 3], and standardized in [1]. The accuracy of its predictions are good, however more surprising is their generality over a variety of acoustic conditions.

Humans frequently perform near Bayes efficiency in psychophysical tasks. Human behavior at tone detection, broadband stimulus detection, and detection or discrimination of more complicated stimuli are all modeled well using Bayes rule and probability distributions of signals in the auditory periphery [4, 7]. Speech recognition involves time-varying signals, higher dimensionality, and more complexity in other ways than the formerly mentioned psychophysical tasks. Humans may not be able to learn all relevant probability distributions and evaluate the most likely classification using Bayes rule when recognizing speech, however they probably do something optimal given the computational limitations of the brain. The feature extraction choices made by a better performing learning machine (the brain) could be helpful for the design of feature extraction for automatic speech recognition.

The Articulation Index model of human performance tells us which factors most affect human speech recognition performance. Some factors, such as sound level and filtering (without changing the speech-to-noise ratio as a function of frequency) have a mild effect on human speech recognition performance while another factor, the speech-to-noise ratio as a function of frequency, has a strong effect on speech recognition performance. We propose that it is worth testing feature extraction techniques which are designed to mimic these properties.

In Section 2 we describe three sets of features for a pattern recognizer: One set is essentially the short-time spectra, another is the short time spectra of speech enhanced using spectral subtraction [10], and the third is informed by the Articulation Index. The Articulation Index-based features mimic a property of human speech perception which is probably learned from speech data by the brain, and consequently exhibit generality over masking noise spectrum and level. The three sets of features will be used to recognize speech in noise for two noise spectra, and at several noise levels. The recognition accuracy provided by these three feature sets is compared in Section 3. The relative advantages and disadvantages of each feature type are discussed in Section 4.

## 2. Methods

Experiments described in this section seek to determine whether a set of features derived from the Articulation Index has any favorable qualities or disadvantages compared to features based on the short-time spectra, with or without spectral subtraction. This section will describe the Articulation Index, three sets of features, the pattern recognizer used in this experiment, hypotheses which will be tested, and the corresponding set of experimental conditions.

### 2.1. Articulation Index

The Articulation Index as described in [6, 5] accounts for the factors of masking, filtering, sound level, and several effects in the human auditory periphery such as neural saturation and upward spread of masking. The most important factors affecting the Articulation Index are masking and filtering, so a simplified version of the Articulation Index documented in [3] is used which accounts only for these effects:

$$AI = \frac{1}{30}\frac{1}{K}\sum_{k}\min(30{,}10\log_{10}(1 + c^2\frac{\sigma^2_{s,k}}{\sigma^2_{n,k}})) \quad (1)$$

The summation index $k$ denotes each of many (denoted $K$, usually 20) adjacent and disjoint bands of the audio spectrum between 200 and 7500 Hz. The frequency edges of these bands prescribed in [6, 5] are derived from perceptual experiments, and are proportional to human auditory frequency resolution [2]. The symbols $\sigma_{s,k}$ and $\sigma_{n,k}$ are the speech and noise root-mean-squared levels in the frequency band corresponding to $k$. The parameter $c$ is described by [6] as “difference ... between the intensity in a critical band exceeded by 1 percent of 1/8 second intervals of received speech and the long [term] average intensity in the same band” which equals 12 dB. The constant $c$ was referred to as the “peak-to-RMS ratio” by [13, 12] which both found that a better prediction of performance could be obtained by using a frequency dependent $c$. The probability of a human incorrectly identifying a phone can be computed from the Articulation Index with

$$P_e = e_{min}^{AI} \quad (2)$$

where $e_{min}$ is a parameter equal to approximately 0.015, which implies a minimum error of 1.5% because the Articulation Index cannot exceed unity.

### 2.2. Short time spectral features

The short-time spectral features (denoted STF) are derived from the set of filters which correspond to each $k$ in Eq. (1). The speech signals are filtered with 15 different filters spanning all frequencies between 200 and 7200 Hz, and have frequency limits [155, 318, 478, 611, 772, 966, 1200, 1481, 1821, 2230, 2724, 3318, 4029, 4881, 5909, 7174] Hz. The output of the filters are narrow-band signals unsuitable for sampling, so the envelope of the filter outputs are extracted by rectification and filtering (the envelope filter has a cutoff frequency of 60 Hz). The resulting envelope of the filtered speech signal for band $k$ is referred to as $s_k(t)$. The signals $s_k(t)$ are sampled once per 17.5 ms. The resulting data rate is 860 dimensions per second (typical automatic speech recognizers use roughly 4000 dimensions per second). The logarithm of the samples of $s_k(t)$ are used because they result in much higher recognition scores.

### 2.3. Articulation Index-based features

The Articulation Index-derived features (denoted AIF) are computed from the short-time spectra features according to

$$a_k(t) = \frac{1}{3}\log_{10}\frac{\mu^2_{n,k} + (s_k(t) - \mu_{n,k})^2}{\mu^2_{n,k}} \quad (3)$$

where $s_k(t)$ is the filter envelope output as before. The symbol $\mu_{n,k}$ is the average of $s_k(t)$ when the input to the system is the masking noise without speech. The signal $a_k(t)$ is sampled at the same rate as the STF (once per 17.5 ms) so the STF and the AIF have the same dimensionality. The noise level $\mu_{n,k}$ is assumed to be known and the speech level is estimated as shown. Equation (3) is an extension of Eq. (1) in time, and all $a_k(t)$ could be used to compute $AI$ as in Eq. (1).

### 2.4. Spectral subtraction features

Spectral subtraction is a technique for denoising speech. The AIF could be viewed as a way of enhancing noisy speech or providing better recognition performance for noisy speech. Spectral subtraction is a possible alternative to the AIF so we compare its performance with the performance of speech enhanced using spectral subtraction, which we will call SSF.

The signal model is

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \quad (4)$$

where $|Y(\omega)|$ is the magnitude spectrum of the noisy speech (and $\angle Y(\omega)$ is its phase). The symbol $|D(\omega)|^2$ represents the known power spectrum of the noise. The symbol $S(\omega)$ is the spectrum of the undistorted speech signal. The known power spectrum of the noise is subtracted from the power spectrum of the noisy speech and the enhanced noisy speech is reconstructed using the phase of the noisy speech. The spectrum of the enhanced version is

$$\hat{S}(\omega) = (|Y(\omega)|^2 - |D(\omega)|^2)^{\frac{1}{2}} e^{j\angle Y(\omega)} \quad (5)$$

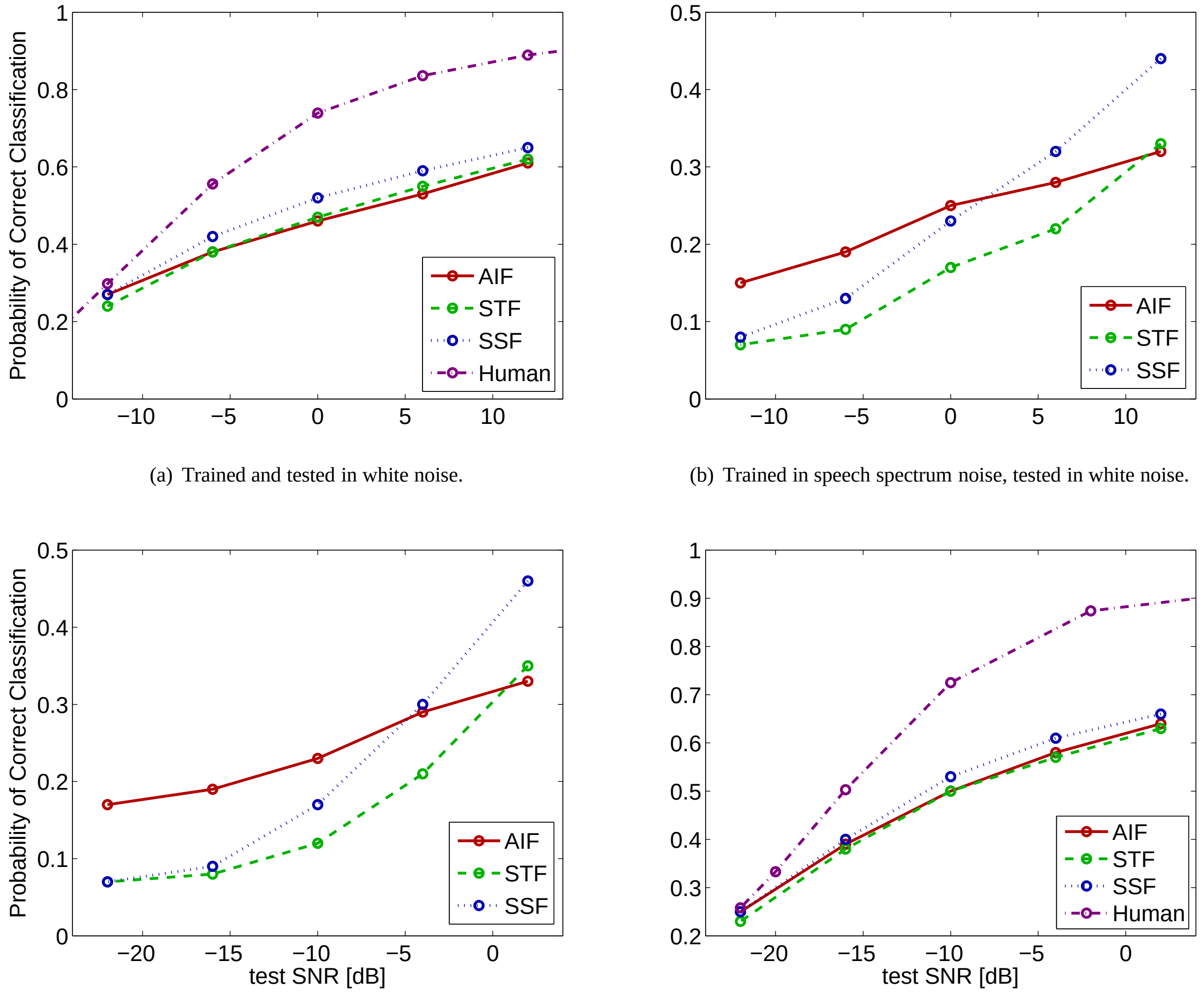
The spectrum of the noisy speech $Y(\omega)$ is obtained by fast Fourier transform. Processing is done in short blocks (in this case 10 ms) so that the noisy phase signal $\angle Y(\omega)$ will have a smaller effect on the quality of the reconstructed speech. The spectrum of the noise $|D(\omega)|$ in this case is known. The enhanced speech sound is generated from $\hat{S}(\omega)$ using the inverse FFT, and features for speech recognition are created by processing the enhanced speech in the way described in Section 2.2.

### 2.5. Pattern recognizer

The pattern recognition task and associated pattern recognizer are chosen so that the task is comparable to one performed by humans in two speech perception experiments [13, 14], so that no assumptions are imposed on the data (e.g. a diagonal covariance matrix), and so the recognizer can plausibly be tested at deeply negative wide-band speech-to-noise ratios (SNRs). In the speech perception experiments described by [13, 14], humans were asked to classify isolated consonant-vowel syllables in two different masking noise spectra at a variety of noise levels. The same task, speech samples, and noise were used in the present study so that machine classifications with the different feature sets could be compared to human classifications.

The recognition task is isolated phone classification for sixteen consonants in CV context. The consonants used are /p,t,k,f,θ,s,ʃ,b,d,g,v,ð,z,ʒ,m,n/. The vowels used are /æ,ʌ,ɛ,ɪ,ʊ,a,e,i,o,u/, although the recognizer does not attempt to classify by vowel. The speech materials are from the “Articulation Index Corpus” (LDC catalog number #LDC2005S22). Approximately 500 examples of each of 16 consonants spoken by twenty people with varying linguistic background are used. The sounds are aligned in time by the “oral release landmark” [15]. Each channel of each token is sampled at nine times, starting 70 ms before the “oral release landmark” and ending 70 ms after. Each token is ultimately represented by a 135 dimensional vector, the product of 15 frequency channels and 9 sample times.

The machine recognition experiment is done with a $K$-nearest neighbors ($K = 9$) classifier, which is used because it is asymptotically Bayes optimal and makes no parametric assumptions



(a) Trained and tested in white noise.

(b) Trained in speech spectrum noise, tested in white noise.

(c) Trained in white noise, tested in speech spectrum noise.

(d) Trained and tested in speech spectrum noise.

Figure 1: Summary of results. Each graph shows the score as a function of test-SNR for a different test/train noise spectrum condition, shown in the caption.

about the structure of the training data. Testing is done by N-fold cross validation. For each token in the data set (7768 tokens in all), the Euclidean distance is computed between that token and all other tokens in the data set. The classification is the most frequently occurring labeled class of the $K$ nearest tokens. Noisy speech samples are generated so they will match the noisy samples used in two speech perception experiments [13, 14]. The masking noise spectrum in [13] is similar to speech, in [14] it is white noise. To evaluate the pattern recognizer in noise, ten noisy versions of each speech sample were generated at each speech-to-noise ratio. The score for a particular token is the average score for the ten noisy versions. Average consonant recognition accuracy without noise using the STF was 74.7%. Human accuracy on the same task (using four vowels and a subset of the speech materials, without noise) was 90.1%.

## 3. Results

The pattern recognizer described above, like most others, is sensitive to mismatch in the noise level or spectrum between the test and training data. Our hypothesis is that the AIF will perform better in cases where the noise level in the test and training set are different, in comparison to the STF and SSF. The recognizer was tested for various mismatches between test and training data using both sets of features. The set of conditions is FEATURE TYPES x TRAINING SNR x TRAINING SPECTRA x TEST SNR x TEST SPECTRA. We use SNRs of [ +12, 6, 0, -6, -12 ] dB in white noise and [ +2 -4 -10 -16 -22 ] dB in speech spectrum noise. The number of conditions is 3 x 5 x 2 x 5 x 2 = 300. Figure 1 shows recognition accuracy for the four test/train noise spectrum conditions as a function of SNR. In the matched spectrum conditions (subfigures (a) and (d)) the testing and training SNR are the same. In the mismatched noise spectrum conditions (subfigures (b) and (c)), the training and test SNRs between speech spectrum and white noise are matched as shown in Table 1. Figure 1 shows 60 of the 300 conditions

| White Noise | 12 | 6 | 0 | -6 | -12 |
|---|---|---|---|---|---|
| Speech Spectrum Noise | 2 | -4 | -10 | -16 | -22 |

Table 1: Wide-band signal-to-noise ratios used for comparisons.

tested.

# 4. Discussion

Figure 1 shows the recognition accuracy provided by the three competing feature types and humans. Each subfigure shows a different testing/training noise spectrum condition. In panels (a) and (d) the test and training noise spectrum are the same, in panels (b) and (c) they are different. Recognition accuracy is lower in the mismatched cases, as expected.

As expected the SSF score is better than the STF score in all conditions, by 1 to 8 percentage points depending on the condition. The advantage of the SSF over the STF is larger in the mismatched noise spectrum conditions, and also larger at higher SNRs.

### 4.1. AIF versus STF and SSF

The first goal of the experiment is to determine whether the AIF have any advantage or disadvantage over more typical features, and specifically whether they generalize better across noise spectrum and level. Subfigures (a) and (d) of Fig. 1 show that the SSF perform best by several percentage points when the testing and training noise spectra are the same, although the advantage disappears as the SNR becomes deeply negative. The STF and the AIF have similar recognition accuracy. Subfigures (b) and (c) show that the AIF performs best over a wide range of SNRs when the testing and training noise spectra are different.

The AIF also have higher recognition accuracy than SSF or STF when the training SNR was low, i.e. SNRs of -6 or -12 dB in white noise, -16 or -22 dB in speech spectrum noise. Not shown in Fig. 1: Compared to the STF, the AIF had equal or greater recognition accuracy in all 40 of these conditions. Compared to the SSF, the AIF had equal or greater recognition accuracy in 33/40 of these conditions.

### 4.2. Observations about human performance

Subfigures (a) and (d) of Fig. 1 compare the percentage of correctly classified phones for the SSF and AIF features used in the current study with human performance on a subset of the same corpus, using the same conditions. In clear recognition accuracy of humans is 15% higher than the machine recognizer, a lead which is maintained at moderate noise levels, and eventually vanishes at deeply negative SNRs.

### 4.3. Comments

The AIF features performed better under conditions of low noise levels or mismatches between test and training data, but not under matched train and test conditions. Future research will seek to isolate the properties that cause the AIF to perform better in these cases, and apply them to automatic speech recognition systems in a way that can provide better performance in varying acoustic conditions.

# 5. References

[1] *Methods for the calculation of the articulation index*, number S3.5. American National Standards Institute, 1969.

[2] J.B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing*, 2(4):567–577, October 1994.

[3] J.B. Allen. Consonant recognition and the articulation index. *J. Acoust. Soc. Am.*, 117(4):2212–2223, April 2005.

[4] N. Durlach, L. Braida, and Y. Ito. Towards a model for discrimination of broadband signals. *J. Acoust. Soc. Am.*, 80:63–72, 1986.

[5] Harvey Fletcher and R.H. Galt. Perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22:89–151, March 1950.

[6] N.R. French and J.C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19:90–119, 1947.

[7] J.J. Hant and A. Alwan. A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise. *Speech Comm.*, 40:291–313, 2003.

[8] K.D. Kryter. Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.*, 34(11):1689–1697, November 1962.

[9] K.D. Kryter. Validation of the articulation index. *J. Acoust. Soc. Am.*, 34(11):1698–1702, November 1962.

[10] Jae S. Lim and Alan V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67:1586–1604, 1979.

[11] H. Müsch. Review and computer implementation of Fletcher and Galt's method of calculating the articulation index. *Acoust. Res. Let. Online*, 2(1):25–30, December 2000.

[12] Chaslav V. Pavlovic and Gerald A. Studebaker. An evaluation of some assumptions underlying the articulation index. *J. Acoust. Soc. Am.*, 75:1606–1612, 1984.

[13] S.A. Phatak and J.B. Allen. Consonant and vowel confusions in speech-weighted noise. *J. Acoust. Soc. Am.*, 121:2312–2326, 2007.

[14] S.A. Phatak, A.W. Lovitt, and J.B. Allen. Consonant confusions in white noise: Effects of noise spectrum and utterance variability. Accepted to *J. Acoust. Soc. Am.*, 2008.

[15] Kenneth N. Stevens. *Acoustic Phonetics*. The MIT Press, Cambridge, Massachusetts, 1998.