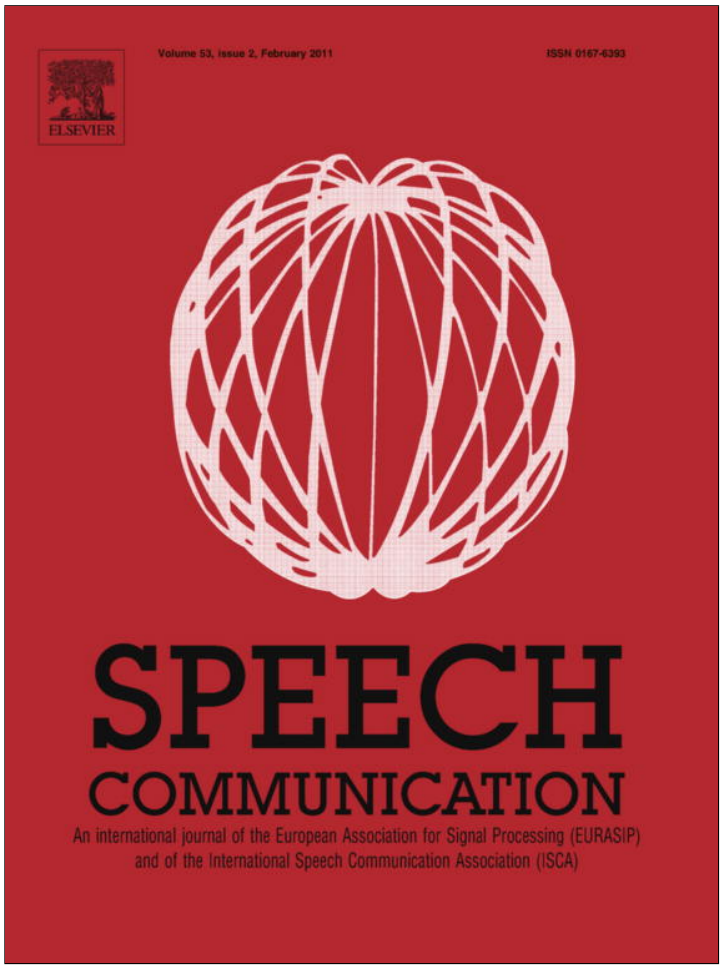

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 185–194
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Intelligibility predictors and neural representation of speech

B.E. Lobdell *, J.B. Allen, M.A. Hasegawa-Johnson

Beckman Institute, 405 N. Mathews Ave., Urbana, IL 61820, USA

Received 28 May 2009; received in revised form 26 July 2010; accepted 30 August 2010

Abstract

Intelligibility predictors tell us a great deal about human speech perception, in particular which acoustic factors strongly effect human behavior, and which do not. A particular intelligibility predictor, the *Articulation Index* (AI), is interesting because it models human behavior in noise, and its form has implications about representation of speech in the brain. Specifically, the Articulation Index implies that a listener pre-consciously estimates the masking noise distribution and uses it to classify time/frequency samples as speech or non-speech. We classify consonants using representations of speech and noise which are consistent with this hypothesis, and determine whether their error rate and error patterns are more or less consistent with human behavior than representations typical of automatic speech recognition systems. The new representations resulted in error patterns more similar to humans in cases where the testing and training data sets do not have the same masking noise spectrum.

Keywords: Speech perception; Articulation Index; Speech recognition; Speech representation

1. Introduction

The authors of (Hermansky, 1998; Allen, 1994) inspire us to (1) examine existing knowledge of human speech perception, (2) employ transformations of speech which simplify the relationship between acoustics and human perception, and (3) use a task which allows machine recognition behavior to be compared in a comprehensible way with human behavior (which is phone classification). The goal of this paper is to examine some qualitative knowledge of human speech perception, and address questions about the structure of the classifier humans use to perform phone transcription.

A great deal of descriptive knowledge exists about speech perception, including:

1. Experiments which find "cues" indicating membership to various phonetic categories by modifying the time-frequency content of a speech signal and observing human classifications. Many phonetic categories have been investigated in this way (Stevens and Blumstein, 1978; Kewley-Port et al., 1983; Cooper et al., 1952; Jongman, 1989; Hedrick and Ohde, 1993; Repp, 1986, 1988; Sharf and Hemeyer, 1972; Darwin and Pearson, 1982).
2. Measurement of human classification accuracy as a function of distortion: removal of fine spectral content, temporal modulations, representation of speech exclusively by formants, etc. (Shannon et al., 1995; Remez et al., 1981; Drullman et al., 1996; Furui, 1986).
3. Models of human behavior as a function of physical qualities of a speech communication channel, such as noise level and filtering. These models of human behavior are called *intelligibility predictors*. The most notable are the Articulation Index (French and Steinberg, 1947) and Speech Transmission Index (Houtgast and Steeneken, 1980).

* Corresponding author. Tel.: +1 217 417 0808.
E-mail addresses: lobdellb@gmail.com (B.E. Lobdell), jontalle@illinois.edu (J.B. Allen), jhasegaw@illinois.edu (M.A. Hasegawa-Johnson).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.016

These studies have contributed greatly to speech and hearing science, audiology, and psychology; however, they arguably have little effect on the design of machine speech recognition systems. This is likely because they describe human behavior, without attempting to infer how they do it. This study is different in that it attempts to infer something about the structure of the human phone classifier.

Systems, known as *intelligibility predictors*, were developed to aid the design of speech communication equipment and auditoria. They are models of human performance as a function of parameters of a speech communication channel. The *Articulation Index* (AI) models the phone error rate as a function of masking noise spectrum and channel filtering. It is described in (French and Steinberg, 1947; Fletcher and Galt, 1950), reformulated in (Kryter, 1962a; Müsch, 2000; Allen, 2005), verified in (Kryter, 1962b; Ronan et al., 2004; Pavlovic and Studebaker, 1984), and standardized in (ANSI, 1969, 1997). The accuracy and generality of its predictions over a variety of acoustic conditions is remarkable.

The AI model of human phone error rate indicates that the most important factor affecting human performance is the speech-to-noise ratio as a function of frequency. The AI is the frequency-average of the non-linearly transformed speech-to-noise ratio (described in detail in Section 1.2). There are numerous modifiers which compensate for sharp filtering, high speech levels, loud maskers, or sharply band-pass maskers, all of which evoke effects in the auditory periphery. It may be deduced from the formulation in (Fletcher and Galt, 1950) that these effects play a relatively small role in typical listening conditions. In fact, another formulation (French and Steinberg, 1947) considers fewer of these peripheral effects presumably because they were not seen as necessary. There is also empirical evidence that a reasonably good prediction of intelligibility can be obtained from an even simpler formulation (Phatak et al., 2008).

It seems reasonable to expect that the human brain keeps a running estimate of prevailing noise and filtering conditions, and uses them to interpret acoustic signals, including speech. This notion was suggested by Hermansky and Morgan (1994), who then developed a representation of speech which ignored the effects of slowly varying filtering and noise conditions. It is also substantiated by Drullman et al. (1994b), which showed that low frequency modulations do not affect human performance. The effectiveness of the Articulation Index has been thought to imply that the brain estimates background noise levels, and only "sees" speech if it is unlikely to have come from the background noise. French and Steinberg (1947) put forth this interpretation:

> When speech, which is constantly fluctuating in intensity, is reproduced at sufficiently low level only the occasional portions of highest intensity will be heard ...
>
> If W is equal to the fraction of the time intervals that speech in a critical band can be heard, it should be possible to derive W from the characteristics of speech and hearing ... it will be appreciated that there are certain consequences that can be tested if the hypothesis is correct that W is equal to the proportion of the intervals of speech in a band which can be heard. There are...

The symbol W is essentially the logarithm of the frequency-specific signal to noise ratio. The intelligibility prediction produced by the AI is essentially the exponent of the average of W over frequency.

They conclude that the speech-derived estimates of W are consistent enough with perceptual data to endorse their hypothesis that W is proportional to the percentage of time intervals during which the speech signal is unlikely to have come from the noise. They use the phrase *can be heard* in a way which seems synonymous with signal detection. Also, an AI model parameter (denoted p in the formulation by French and Steinberg (1947)) is specifically related to the probability distribution of speech (the level in decibels which is higher than 99% of speech levels), and is employed in a way which assumes speech is detectable if its level is greater than a threshold. Two studies (Phatak and Allen, 2007; Pavlovic and Studebaker, 1984) have shown that frequency-specific values for p based on the level distribution of speech offer a better prediction of human recognition accuracy, supporting this view. The meaning of W, the form of the AI prediction of intelligibility, and its relationship to signal detection will be discussed in more detail in Section 1.2.

The AI predicts average phone error rate for a large amount of phonetically balanced speech, based on the average spectrum of speech, and information about the acoustic conditions. The interpretation of the AI offered above is based on the average spectrum of speech and average phone error rate. In this paper we will attempt to determine whether this interpretation holds for classification of individual utterances, based on the acoustics of individual utterances.

1.1. Parameterization of the speech signal

The following paragraphs place this study in the context of research on machine speech recognition, and human speech perception.

Standard representations of speech for speech recognition include the mel-frequency cepstral coefficients (MFCCs) and perceptual LPC (PLP). Davis and Mermelstein (1980) demonstrated that warping the frequency axis to a perceptually-based scale improves word discriminability. Hermansky (1990) demonstrated that an all-pole summary of the loudness spectrum (PLP) exhibits less inter-speaker variability than the raw loudness spectrum. Optimization-based approaches have been adopted recently; for example, transforming the speech signal to maximize information content (Padmanabhan and Dharanipragada,

2005), or transforming the speech signal into a form which can be parsimoniously represented by parametric distributions used in speech recognition systems (Omar and Hasegawa-Johnson, 2004). None of them have supplanted the MFCCs as the dominant representation of speech for automatic speech recognition. The current study is different because it seeks to determine whether a representation of speech in noise is more or less consistent with human behavior, rather than deriving one more appropriate for speech recognition systems.

The idea of using representations of speech inspired by the human auditory system is not new. For example, Hermansky (1990) suggests a representation of speech based on human auditory tuning, level normalization, and compression (which are present in the auditory system). In (Strope and Alwan, 1997) the authors simulate the dynamic activity of the auditory system to emulate a phenomenon called forward masking, and showed that a recognizer based on it is more robust to background noise than conventional systems. Another representation of speech called RASTA (Hermansky and Morgan, 1994), is predicated on an assumption that the brain keeps a running estimate of noise and filtering conditions, and uses them when recognizing speech. All systems compared in the current study use an auditory-like representation of speech similar to PLP (described in Hermansky, 1990). Our intention is to test a particular representation of speech in noise to deduce the structure used to classify phones, rather than test the merits of auditory-like representations of speech, which we already consider to be important.

Studies about representations of speech in noise, and models for detection of speech in noise are especially relevant. Experiments have been done (for example in Viemeister and Wakefield, 1991; Durlach et al., 1986) which demonstrate that Bayes' rule applied to the probability distribution of auditory signals can predict human performance for some psycho-physical tasks. Hant and Alwan (2003) show that a similar model also predicts discrimination of some speech sounds. This paper is meant to expand the domain of tasks which Bayes' rule can explain.

1.2. The Articulation Index

The Articulation Index models human phone recognition accuracy as a function of filtering and masking conditions. Several versions of the AI (mentioned in Section 1) have been published, which vary in sophistication and correspondingly, their accuracy and convenience. For the sake of brevity, we will describe the version published in (Allen, 2005) which has good accuracy in typical listening conditions.

First, speech is filtered into (in this formulation, 30) disjoint frequency bands with bandpass filters. The edges of these bands were chosen to fit empirical data, and are roughly proportional to the critical bandwidth (Fletcher, 1938; Allen, 1994). The second step is measurement of the speech and noise root-mean-squared levels in each bank, denoted here by $\sigma_{s,k}$ and $\sigma_{n,k}$, respectively, where k indexes the frequency band.

The Articulation Index is computed from

$$\mathrm{AI} = \frac{1}{30}\frac{1}{K}\sum_{k} \min\left(30, 10\log_{10}\left(1 + c^2\frac{\sigma_{s,k}^2}{\sigma_{n,k}^2}\right)\right). \quad (1)$$

The parameter p (and $c = 10^{p/20}$) is related to the quote by French and Steinberg (1947) in Section 1. They describe p as the "difference in db between the intensity in a critical band exceeded by 1% of $\frac{1}{8}$th second intervals of received speech and the long average intensity in the same band," depicted in Fig. 1. Thus p (and c) are related to the threshold which is thought to determine whether humans "hear" speech at a particular frequency and time. French and Steinberg (1947) use A in place of our symbol AI and represent the argument of the summation in Eq. (1) with W. They hypothesize that W "is equal to the fraction of the time intervals that speech in a critical band can be heard" which, in terms of Fig. 1, suggests some level on the abscissa which represents a threshold: speech intervals above the threshold can be heard and those below the threshold cannot. They suggest W could be computed by integrating the speech probability distribution in Fig. 1 above this threshold.

The probability of a human incorrectly identifying a phone can be computed from the Articulation Index (Eq. (1)) with

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

where e_{min} is a parameter equal to approximately 0.015.

1.3. Hypothesis statement

The experiment described in this paper is meant to test the hypothesis that humans' phone transcriptions for an acoustic waveform are based on the time-frequency signal-to-noise ratio rather than the short-time spectral level:

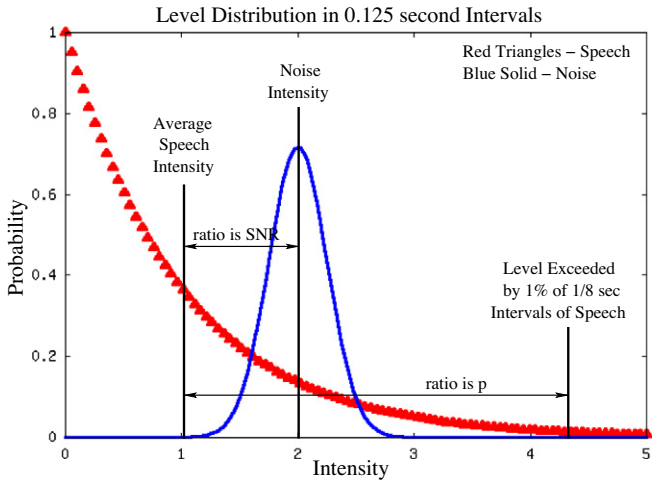

Fig. 1. This cartoon shows the probability distribution of speech and noise levels in 0.125 s intervals. It also illustrates the AI model parameter p from (French and Steinberg, 1947) in relation to the speech and noise level distributions.

A particular time-frequency sample will affect classification only if that sample is unlikely to have resulted from the prevailing noise level in that spectral channel. This is a difficult proposition to test directly because many samples interact with each other in the brain, and our perceptual experiments are not sensitive enough to measure the effect of a single sample. Rather than attempt to directly test this hypothesis, which in our view is ill advised, we will classify speech sounds with several representations of speech, and examine the results to see which are most consistent with human classifications.

Four representations of speech will be tested:

1. The power spectrum. Many automatic speech recognition systems observe a linear transform of the log power spectrum (mel-frequency cepstral coefficients), therefore the power spectrum maybe considered analogous to those usually used in speech recognition. These will be referred to as the STF (spectro-temporal features).
2. A representation based on the Articulation Index, which is essentially the speech-to-noise ratio as a function of time and frequency. This will be referred to as the AIF.
3. A thresholded version of the AIF. A particular time-frequency pixel is unity if its SNR is greater than some threshold, and zero otherwise. These will be referred to as the AIBINF ("BIN" for binary).
4. A version of the STF enhanced by spectral subtraction, which is called SSF.

This study will evaluate these speech representations based on the similarity between the mistakes they produce and the mistakes produced by humans in the same acoustic conditions. Greater consistency between human and machine errors is interpreted to mean greater similarity between the human recognition process and the classifier implied by a particular representation. The recognition accuracy provided by the various feature types will be compared, since a better performing feature type will be of interest to speech recognition researchers.

If the AIF leads to mistakes similar to those made by humans, it will support our hypothesis that humans estimate the prevailing noise spectrum and represent speech as an SNR spectrum rather than as the power spectrum of the noisy signal (as in the STF). The drop in performance between AIF and AIBINF determines how much information is gained by using a high level-resolution representation of the signal (as in the AIF) rather than only a single bit (1 = detected, 0 = not detected) for each time/frequency pixel. The SSF are included because the AIF features will be of less engineering interest if they do not provide an advantage over the simple and ubiquitous noise removal technique called spectral subtraction.

1.4. Synopsis

Section 2 describes the speech representations used to test our hypotheses, the human speech classification experiments, the machine speech classification experiments, and the metric used to compare results from them. Section 3 shows the recognition accuracies for each experiment, the relative performance of the various feature types, the similarity between the human and machine mistakes, and the most evident conclusions. The final section discusses their implications to the hypothesis presented above.

2. Methods

2.1. Speech materials

The stimuli used in this study are consonant-vowel sounds from the "Articulation Index Corpus" published by the Linguistic Data Consortium (Catalog #LDC2005S 22). The sixteen consonants [/p, t, k, f, θ, s, ʃ, b, d, g, v, ð, z, ʒ, m, n/] are paired with vowels in all experiments. The average duration of the speech sounds is 500 ms.

The machine experiment uses the sixteen consonants paired with ten vowels, and approximately fifty examples of each consonant-vowel pair. The total number of sounds is approximately 16 consonants × 10 vowels × 50 tokens = 8000 speech sounds. The vowels used were [/æ, ʌ, ɛ, ɪ, ʊ, ɑ, e, i, o, u/].

The human experiments use a smaller number of vowels to limit the experiment time. The experiment with speech spectrum noise paired each of the sixteen consonants with four vowels, which were [/ɑ, ɛ, ɪ, æ/]. There were fourteen examples of each consonant–vowel pair. The total number of sounds is 16 consonants × 4 vowels × 14 tokens = 896 speech sounds. The experiment with white spectrum noise paired each of the sixteen consonants with a single vowel, which was [/ɑ/]. There were 18 examples of each consonant–vowel. The total number of sounds is 16 consonants × 1 vowel × 18 tokens = 288 speech sounds.

2.2. Speech representations

The speech signal is analyzed by a filter bank consisting of 15 filters having frequency limits [155, 318, 478, 611, 772, 966, 1200, 1481, 1821, 2230, 2724, 3318, 4029, 4881, 5909, 7174] Hz. The filters are 5th order elliptical, with ripple of 2% and stop band suppression of 60 dB. 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 Drullman et al., 1994a). The resulting envelope of the filtered speech signal for band k is $s_k(t)$. The speech sounds are manually aligned by the oral release, and sampled 70 ms before the release to 70 ms after the release at intervals of 17.5 ms. The resulting data rate is 15 frequency channels × 1/0.0175 samples per second = 857.14 dimensions per second. Fifteen filters and nine sample times (−70 ms to +70 ms, every 17.5 ms) provides a 135-dimensional representation of each speech token.

2.2.1. Baseline features

The baseline representation (denoted STF) is a non-whitened version the MFCCs typically used in automatic speech recognition. The symbol $s_k(t)$ refers to the envelope of the signal proceeding from the kth filter. The 135-dimensional representation of a speech token is the logarithm of a sampled version of $s_k(t)$, which is $\log s_k(0.0175n)$ for $k = 0,\ldots 14$, and $n = -4,\ldots 4$.

2.2.2. Articulation Index-based features

The Articulation Index-derived representation (denoted AIF) is computed from $s_k(t)$ and estimates of the noise level $E[n_k^2(t)]$ and $E[n_k(t)]$ with

$$a_k(t) = \log_{10} \frac{E[n_k^2(t)] + (s_k(t) - E[n_k(t)])^2}{E[n_k^2(t)]}. \quad (3)$$

The symbols $E[n_k(t)]$ and $E[n_k^2(t)]$ are the time-average of $s_k(t)$ and $s_k^2(t)$, respectively, when the input to the system is the masking noise without speech. Eq. (3) is not identically equivalent to a time-unrolled version of the Articulation Index but is a close approximation which has the crucial properties that (1) it fluctuates randomly around 0 when $s_k(t)$ is only noise, (2) as the speech-to-noise ratio grows, the function approaches Eq. (1), and (3) most importantly $a_k(t)$ is small for segments of noisy speech which are likely to contain only noise. Item (3) is the property of the Articulation Index formula which French and Steinberg (1947) hypothesize explains its predictive power.

The representation used for machine classification in this study is a sampled version of $a_k(t)$, which is $a_k(0.0175n)$ for $k = 0,\ldots 14$, and $n = -4,\ldots 4$. This representation is denoted AIF.

2.2.3. Binary AI-based features

This representation (denoted AIBINF) is a thresholded version of the AIF. The value is 1 if Eq. (3) is greater than 0.3, and 0 otherwise. This can be interpreted to mean that a pixel will be labeled "1" if there is greater than 99.9% chance that the observed level was speech and not noise.

2.2.4. Spectral subtraction features

Spectral subtraction is a technique that can be used to mitigate the effects of noise on speech recognition (Boll, 1976).

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

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

The spectral subtraction technique is implemented via the analysis-modification-synthesis methodology. The spectrum of the noisy speech $Y(\omega)$ is obtained by fast Fourier transform in small blocks, which mitigates the effect of the noisy phase signal $\angle Y(\omega)$. The block size used in this study is 10 ms. The speech is modified in the spectral domain according to Eq. (5). The enhanced speech $\hat{s}_k(t)$ is synthesized from $\widehat{S}(\omega)$ using the inverse FFT. The representation used for machine classification in this study is a sampled version of $\hat{s}_k(t)$, which is $\hat{s}_k(0.0175n)$ for $k = 0,\ldots 14$, and $n = -4,\ldots 4$. This representation is denoted SSF.

2.3. Human perception experiments

Machine classifications are compared with human classifications in an analogous experiment: classification of isolated consonant–vowel syllables. Human classifications were collected in two experiments conducted at the University of Illinois. All test subjects had normal hearing and were from the University of Illinois community. The experiment was administered by an automatic computer program which tabulated the listeners' classifications of the speech materials. The listeners heard the sounds over Sennheiser HD265 headphones, generated by a "Soundblaster Live!" sound card, inside an Acoustic System model 27930 anechoic chamber. The experiment in white noise involved the 16 consonants paired with the vowel /ɑ/ at [Clear, 12, 6, 0, −6, −12, −15, −18] dB speech-to-noise ratio. The experiment in speech spectrum noise involved consonants paired with the vowels /ɑ, ɛ, ɪ, æ/ at [Clear, −2, −10, −16, −20, −22] dB speech-to-noise ratio. Response probabilities for these experiments were calculated for each talker, consonant, vowel, and speech-to-noise ratio condition. More details about these experiments can be found in (Phatak and Allen, 2007; Phatak et al., 2008).

2.4. Automatic speech classification experiment

An asymptotically Bayes optimal pattern recognizer is used to (1) avoid assumptions about the statistics of the features (e.g. that the dimensions are uncorrelated, or that they conform to some parametric distribution) and (2) achieve above-chance recognition accuracy at the deeply negative speech-to-noise ratios used in the human experiments.

The speech sounds described in Section 2.1 are classified by a K-nearest neighbors based classifier ($K = 9$), using the four representations of speech described in Section 2.2. The speech sounds were from the same corpus as in the human experiment, but involved more examples of each consonant. The output of the classifier was a consonant label. Each of the 16 consonants was exemplified by approximately 500 utterances, with a total of 7768 (approximately 16×500) token for all consonants. Consonants were

classified using 7768-fold cross validation: Each token was classified by computing the Euclidean distance in the 135 dimensional feature space between itself and each other token; the assigned class was the most frequently occurring class among its K closest neighbors. Noisy versions of each sound were created, which had the same noise level and spectrum as those used in the human experiments. Ten noisy realizations were mixed with each token and classified to generate the response probabilities in noise.

Robustness to a variety of acoustic conditions is an outstanding quality of speech recognition by humans. The classifier will classify the sounds with various mismatches between the testing and training conditions to contrast the behavior of the various systems with human behavior in this important case. The sounds are classified in white noise at [12, 6, 0, −6, −12] dB SNR, and in speech spectrum noise at [2, −4, −10, −16, −22] dB SNR. Each possible mismatch will be tested, so there will be {4 feature types} × {2 test noise spectrum} × {5 test SNR} × {2 training noise spectra} × {5 training SNR} = 400 conditions.

2.5. Comparison metric for response probabilities

A symmetrized Kullback–Liebler (KL) metric is used to compare human and machine response probability distributions. The KL-metric is a measurement of the difference between a pair of probability distributions. It has the unit of *bits* (if the logarithm has base 2) and its minimum value is zero when the distributions being compared are identical (Cover and Thomas, 2006).

The comparison metric should be symmetric (as we have no basis on which to order its arguments), and it should be finite even if one probability mass function has outcomes with zero probability and the other does not. For that reason we will use the following adaptation of the KL-metric, which is symmetric and always finite.

$$D(p,q) = \frac{1}{2}\left(\sum_i p_i \log \frac{p_i}{p_i/2 + q_i/2} + \sum_j q_j \log \frac{q_j}{p_j/2 + q_j/2}\right), \tag{6}$$

where p_i and q_i are probability mass functions for human or machine classifications. The symbol i indexes the response options /p, t, k, f, θ, s, ʃ, b, d, g, v, ð, z, ʒ, m, n/. The arguments p_i and q_i are a particular row j of the confusion matrices being compared $P_{ij,human}$ and $P_{ij,machine}$. P_{ij} is the probability that symbol i was classified as symbol j. Section 3 will compare $P_{ij,human}$ and $P_{ij,machine}$.

3. Results

In this section we will summarize the results of the experiment: recognition accuracy, and similarity to human response patterns. The first is relevant to evaluating the AI-based features' value for automatic speech recognition, and the second to our hypothesis about human speech perception.

Fig. 2 shows data from a subset of the conditions. Panels (a) and (d) of Fig. 2 show recognition accuracies for the conditions where the test noise spectrum and level match the training noise spectrum and level. The SSF provides the best recognition accuracy in these conditions, exceeding the recognition accuracy of the AIF and STF by a few percent. The AIBINF has markedly lower performance in these conditions. Panels (b) and (c) of Fig. 2 show cases where the test and training noise spectrum are different. In these cases, the AIF or AIBINF provide the best performance at the 3–4 lowest speech-to-noise ratios, while the SSF provide best performance at the highest 1–2 speech-to-noise ratios. Human recognition accuracy is substantially greater than all the machine systems. Humans did not need to be trained to perform well in either white noise or speech-shaped noise; human speech recognition accuracies are plotted only in comparison to the "matched" automatic classification experiments (subplots (a) and (d)). The recognizer correctly classified 74.7% of the speech sounds in clear using the STF (SSF is equivalent to STF, and AIF, AIBINF are undefined in the absence of noise), while humans achieved 90% in clear.

Table 1 shows recognition accuracy results for all 400 conditions in the machine experiment. The column indicates which noise condition was used to train the classifier, the row indicates which noise condition was classified. Each cell of the table contains four numbers, which show the recognition accuracy in percentage arranged as follows:

SSF/STF AI/AIBINF

Fig. 3 shows which representations achieved the highest recognition accuracy in each condition shown in Table 1. A red cell (with "x") indicates that SSF had the highest recognition accuracy, blue (with "^") indicates AIF, and yellow (with "*") if AIBINF had the highest accuracy. The baseline features STF never had the highest recognition accuracy. The SSF usually had the highest recognition accuracy when the testing and training noise spectra were the same (upper-left and lower-right quadrants), although the AIF and AIBINF were often better in cases where there was a large mismatch in noise level. The AIF or AIBINF usually had a much larger advantage when they were best, than did SSF or STF.

Table 2 compares human and machine error patterns for each condition using the metric describe in Section 2.5. Each cell of Table 2 and approximately fifty examples of each consonant–vowel Eq. (7):

$$\frac{1}{16}\sum_{i=1}^{16} D\left(P_{ij,human}, P_{ik,machine}\right). \tag{7}$$

There are 400 machine confusion matrices $P_{ij,machine}$, resulting from 100 noise testing and training conditions, and the four systems described in Section 2.2. The human probability distribution $P_{ij,human}$ is the same in each row of Table 2.

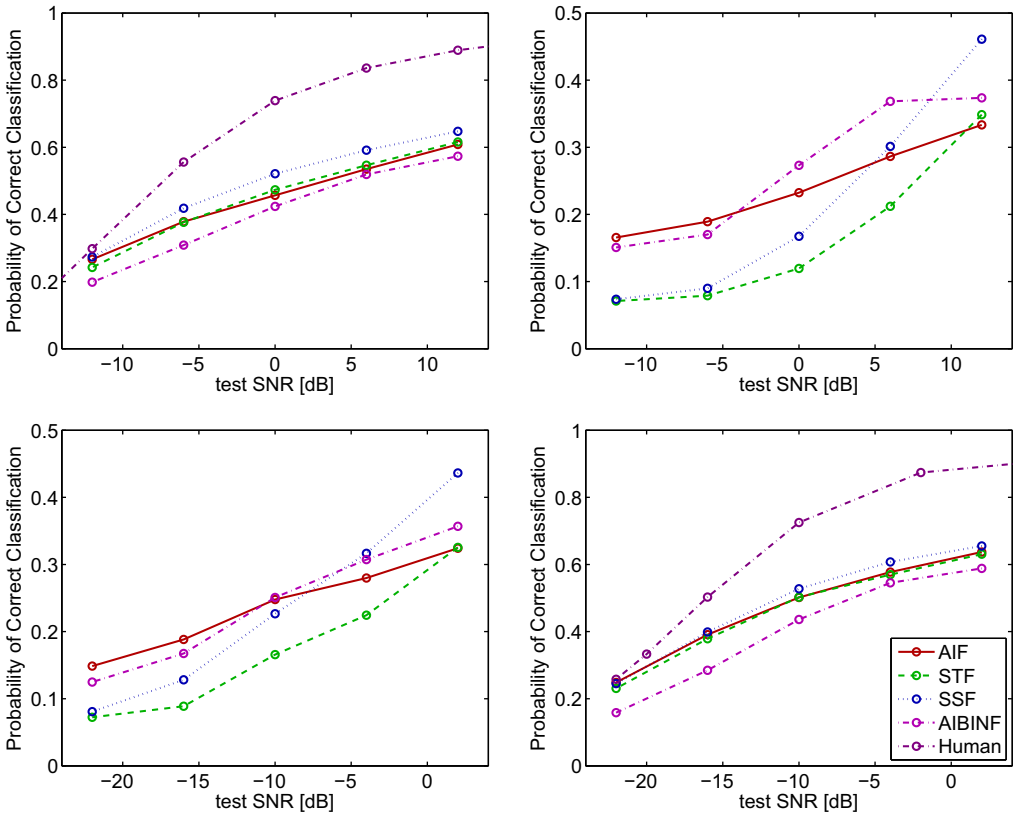

Fig. 2. This figure shows the recognition accuracies for the four systems tested in the machine experiment, and the human experiment. The upper-left and lower-right plots show the case where the testing and training noise spectra are same. The upper-right and lower-left plot show cases where the testing and training noise spectra are mismatched. In the same-noise spectra conditions, the recognition accuracies are approximately equal, with a slight advantage for the SSF. In the mismatched conditions, the AIF or AIBINF have an advantage.

Fig. 4 shows which representation has the smallest distance to human error patterns. The STF usually has the smallest distance (most similar) to human error patterns when the testing/training data were matched, however in mismatched cases the AIF or AIBINF usually have the smallest distance.

4. Discussion

4.1. Review of hypotheses

Humans do not suffer from train-test mismatch in speech classification problems, but machines do. An automatic classifier using AI-based features suffers less than a classifier using spectral subtraction: its classification accuracy is higher, and its confusion matrices more closely resemble the confusion matrices produced by human subjects (lower symmetrized KL divergence).

Classification accuracy and KL divergence are correlated in only two respects: they are degraded by train-test mismatch, and the degradation is usually reduced by use of AIF or AIBINF. In many cases, classification accuracy is a poor predictor of KL divergence and vice verse.

4.2. Value for engineering problems

The AI-based features were not far from the performance of SSF in matched cases, and exceeded the performance of SSF in mismatched cases. Automatic speech recognizers are often brittle to changes in acoustic conditions, making a representation robust to such changes valuable. The feature types tested here are not yet well suited for a practical recognizer because dimensions of the feature space will be correlated, and cannot be represented efficiently by parametric probability distribution models used in speech recognizers. We also have not provided a means to estimate the noise level, or investigated how the brain measures the noise spectrum, although such methods are available (Martin, 2001; Lee and Hasegawa-Johnson, 2007).

The AIBINF features performed best in some conditions, and never much less that the best performing features even though their data rate is a small fraction of the data rate of the other feature-types tested. This could be useful in that it would greatly reduce the training data requirements, and simplify the acoustic model, as each state emission probability distribution is now (multivariate) Bernoulli rather than multivariate Gaussian.

4.3. Limitations

There are some issues with the experiment which bear mentioning. The K-nearest neighbors recognizer was used because it is asymptotically Bayes optimal (i.e., with an infinite number of training examples). However the number

Table 1
This table shows recognition accuracies from the machine experiment. Each cell shows the results for a particular (testing noise spectrum and SNR) × (training noise spectrum and SNR) condition. Each cell contains four numbers, which correspond to each of the four feature types. Upper-left is SSF, upper-right is STF, lower-left is AIF, lower-right is AIBINF.

Test SNR	Train SNR									
	12	6	0	−6	−12	2	−4	−10	−16	−22
12	65/62 61/57	53/46 40/44	24/16 15/21	9/8 3/7	6/6 4/4	46/35 33/37	28/19 23/31	15/10 15/19	9/9 9/9	8/7 7/6
6	58/54 52/49	59/55 53/52	43/34 31/35	13/11 11/16	7/7 4/6	45/36 37/36	30/21 29/37	16/13 20/25	10/9 11/11	8/8 8/7
0	43/35 33/33	52/45 43/44	52/47 46/42	30/23 22/24	9/8 9/8	37/26 37/35	31/22 35/39	17/12 23/27	10/8 17/12	8/7 11/9
−6	25/18 18/12	31/24 24/22	41/34 34/33	42/38 38/31	18/15 16/15	22/14 28/22	21/13 34/34	16/11 31/30	9/8 19/17	7/7 15/14
−12	11/9 14/11	15/11 17/15	19/15 21/19	28/23 27/24	27/24 27/20	10/7 21/16	11/7 27/22	10/8 31/26	9/9 24/19	7/7 17/15
2	44/33 32/36	32/23 26/29	22/16 19/20	13/10 12/15	9/8 6/7	66/63 64/59	52/46 42/45	22/17 18/20	10/9 9/10	7/7 7/7
−4	41/31 31/30	32/22 28/31	23/17 24/24	16/11 19/19	10/9 10/10	60/55 55/50	61/57 58/55	40/33 32/34	14/11 13/15	8/8 8/9
−10	33/26 32/25	29/24 29/28	23/17 25/25	16/11 22/23	10/8 17/17	44/38 37/34	52/47 49/46	53/50 50/44	26/22 22/21	9/8 9/11
−16	22/17 25/20	24/17 25/23	21/14 21/20	13/9 19/17	8/6 19/16	28/22 29/21	35/28 35/29	40/37 42/35	40/38 39/28	17/15 17/15
−22	14/12 26/23	16/11 26/24	15/11 23/21	12/10 18/17	8/7 15/12	17/14 20/19	21/17 23/24	25/22 29/27	29/27 33/24	25/23 25/16

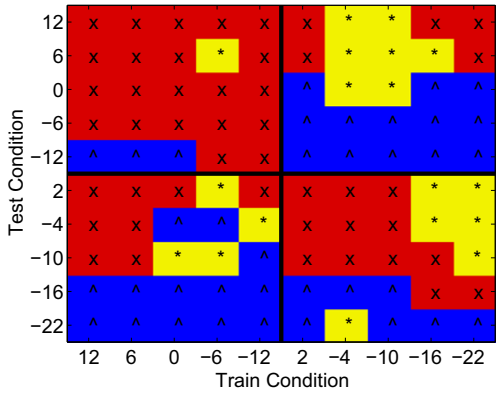

Fig. 3. This is a graphical representation of the data in Table 1 showing which system had the highest recognition accuracy in each testing/training condition. A cell is red (with "x") if SSF had the highest recognition accuracy, blue (with "^") if AIF had the highest accuracy, and yellow (with "*") if AIBINF had the highest accuracy. The baseline features STF never had the highest recognition accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

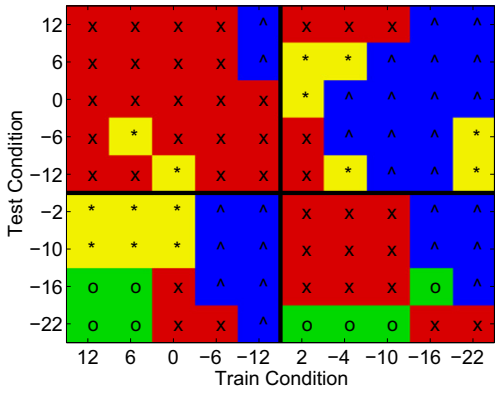

Fig. 4. This is a graphical representation of the data in Table 2 showing which system had the smallest KL-metric from human behavior in each testing/training condition. A cell is red (with "x") if SSF had the lowest KL-metric, green (with "o") if STF had the lowest KL-metric, the blue (with "^") if AIF had the lowest KL-metric, and yellow (with "*") if AIBINF had the lowest KL-metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of training examples used in this experiment is finite, so the relative performance of the four systems tested might be different if there were more training data. The parametrization of speech will also effect the relative performance of each response alternative. Another smaller experiment was conducted to relieve these concerns. It involved recognizing three categories, using three different classifiers: K-nearest neighbors, multi-layer perceptron, and Gaussian PDF. The stop consonants /p, t, k, b, d, g/ were categorized by place of articulation [bilabial, alveolar, velar]. The speech sounds were represented only by the spectrum of the release burst for the stop consonants (15 spectral slices × 2 time slices), rather than by the 135 dimensions used in the primary experiment. The neural network had 30 input nodes and 17 hidden nodes. The Gaussian PDF model has 30 dimensions. The relative classification

Table 2
This table shows the KL-metric between machine classifications and human classifications in the same test condition. Each cell shows the results for a particular testing/training noise spectrum/SNR condition. Each cell contains four numbers, which are for each of the four feature types. Upper-left is SSF, upper-right is STF, lower-left is AIF, lower-right is AIBINF.

Test SNR	Train SNR									
	12	6	0	−6	−12	2	−4	−10	−16	−22
12	.12/.13 .14/.17	.16/.19 .20/.25	.27/.34 .36/.42	.45/.54 .55/.70	.68/.75 .66/.72	.31/.43 .38/.36	.31/.41 .40/.43	.38/.47 .39/.51	.52/.61 .50/.54	.68/.74 .50/.58
6	.16/.20 .25/.23	.12/.14 .15/.17	.17/.21 .22/.23	.34/.41 .41/.47	.57/.65 .55/.62	.42/.51 .43/.38	.41/.53 .37/.36	.41/.48 .37/.42	.48/.60 .47/.47	.61/.73 .47/.51
0	.39/.49 .52/.40	.18/.23 .27/.24	.12/.14 .15/.18	.19/.24 .23/.24	.41/.48 .41/.47	.47/.56 .49/.46	.47/.56 .39/.39	.49/.57 .34/.35	.49/.61 .45/.47	.61/.72 .43/.46
−6	.62/.67 .76/.64	.45/.48 .53/.42	.19/.24 .27/.26	.12/.12 .13/.19	.21/.24 .23/.28	.53/.61 .58/.54	.50/.58 .39/.42	.49/.61 .34/.34	.59/.71 .40/.45	.60/.66 .39/.39
−12	.64/.65 .78/.77	.59/.61 .70/.60	.45/.48 .56/.44	.20/.23 .25/.26	.11/.11 .13/.20	.54/.56 .76/.73	.54/.59 .54/.53	.52/.60 .33/.34	.65/.72 .29/.33	.68/.64 .33/.33
−2	.41/.52 .55/.40	.38/.49 .40/.34	.38/.48 .33/.31	.49/.61 .38/.40	.66/.74 .47/.55	.15/.17 .19/.20	.12/.14 .13/.15	.22/.25 .24/.25	.39/.46 .38/.44	.58/.63 .53/.58
−10	.54/.59 .65/.52	.52/.57 .61/.44	.53/.61 .47/.38	.49/.61 .29/.31	.57/.64 .29/.32	.43/.52 .59/.48	.18/.24 .25/.25	.09/.10 .10/.14	.17/.20 .17/.21	.36/.40 .30/.34
−16	.56/.52 .81/.77	.49/.49 .61/.60	.51/.53 .58/.58	.59/.65 .49/.52	.57/.59 .27/.43	.57/.61 .71/.67	.45/.48 .58/.60	.23/.27 .31/.39	.09/.09 .09/.19	.17/.18 .14/.27
−22	.60/.51 .86/.87	.45/.44 .74/.74	.46/.48 .66/.66	.52/.55 .58/.64	.59/.62 .50/.69	.70/.68 .82/.82	.63/.62 .78/.76	.53/.53 .70/.71	.31/.35 .43/.61	.12/.13 .15/.55

accuracy of the AIF and STF were roughly the same as in the primary experiment, and the performance of the three classifier systems were also roughly the same. The similarity of performance between three types of classifiers, and consistency of results with the first experiment relieves our concern that the result in the first experiment is a fluke (see Table 3).

It was suggested in one of the original papers about the Articulation Index (Fletcher and Galt, 1950) that the human behavior for some speech sounds is best modeled by a function of the speech-to-noise ratio (denoted *R-M*) while some other sounds are better modeled by a function of the speech spectrum (denoted *R*):

> There have been two points of view advanced as to how an observer interprets the speech sounds in the presence of a noise. The first point of view assumes that the relative position of the speech components with respect to the threshold in the noise determines the factor *F* in Eq. (19). According to this point of view the effective response has been lowered by the threshold shift *M* due to the noise, so that the quantity R–M takes the place of *R* in determining the factor *F*. The second point of view, which was taken by one of the present authors in an earlier formulation of this theory, assumes that the level of the speech components with respect to each other is the principle influence in determining *F*.
> The articulation tests indicate that some of the sounds of speech act in accordance with the first assumption, while the other sounds follow the second assumption.

Table 3
This table summarizes results from a smaller classification experiment, the purpose of which was to demonstrate that the results of the larger experiment in this paper were not an artifact of the classifier used. Each cell shows the difference $P_c(AIF) - P_c(STF)$ averaged over three classifiers (which were MLP, Gaussian PDF-based, and K-nearest neighbors). The second number in each cell (right of "/") shows the standard deviation over the three classifiers. In all cases the standard deviation over classifiers is relatively small. In the matched cases, the difference between AIF and STF performance is not significant, and the mismatched case it is, a result consistent with the larger experiment discussed in this paper.

SNR	High	Mid	Low
Matched	1.4/3.3%	7.6/9.5%	2.5/3.8%
Unmatched	8.4/2.0%	11.2/2.8%	11.1/1.7%

The experiment described in this paper does not attempt to test the statement in (Fletcher and Galt, 1950), so the effect they describe may cloud our result. If their hypothesis is correct, better recognition accuracy (and similarity to human error patterns) could be achieved by recognizing each consonant (or sub-phone unit) with the most appropriate feature-type (AIF or STF). However, we do not have *a priori* knowledge of which system to use on each consonant, and some hard to justify choices would have to be made about how to combine information from each of the recognizer systems.

5. Conclusions

We classified speech sounds with several representations of speech meant to help us determine which representation is more consistent with human behavior.

The AI-based representations performed better and had error patterns more consistent with humans in cases where the testing and training noise spectrum or level were mismatched. This property could be valuable in a practical recognizer because robustness to changes in conditions is a major problem in speech recognition.

A thresholded version of the AI-based features did surprisingly well, and may indicate that precise representation of the spectrum level is not particularly relevant for the task.

References

Allen, J., 1994. How do humans process and recognize speech? IEEE Trans. Speech Audio Process. 2, 567–577.

Allen, J., 2005. Consonant recognition and the articulation index. J. Acoust. Soc. Amer. 117, 2212–2223.

ANSI (1969). Methods for the calculation of the articulation index, ANSI S3.5.

ANSI (1997). Methods for the calculation of the speech intelligibility index, ANSI S3.5.

Boll, S.F., 1976. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Cooper, F., Delattre, P., Liberman, A., Borst, J., Gerstman, L., 1952. Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Amer. 24, 597–606.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. Wiley and Sons, Hoboken, NJ.

Darwin, C., Pearson, M., 1982. What tells us when voicing has started? Speech Comm. 1, 29–44.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366.

Drullman, R., Feste, J., Plomp, R., 1994a. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95, 1053–1064.

Drullman, R., Festen, J., Plomp, R., 1994b. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Amer. 95, 2670–2680.

Drullman, R., Festen, J., Houtgast, T., 1996. Effect of temporal modulation reduction on spectral contrasts in speech. J. Acoust. Soc. Amer. 99, 2358–2365.

Durlach, N., Braida, L., Ito, Y., 1986. Towards a model for discrimination of broadband signals. J. Acoust. Soc. Amer. 80, 63–72.

Fletcher, H., 1938. Loudness, masking and their relation to the hearing process and the problem of noise measurement. J. Acoust. Soc. Amer. 9, 275–293.

Fletcher, H., Galt, R., 1950. Perception of speech and its relation to telephony. J. Acoust. Soc. Amer. 22, 89–151.

French, N., Steinberg, J., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Amer. 19, 90–119.

Furui, S., 1986. On the role of spectral transition for speech perception. J. Acoust. Soc. Amer. 80, 1016–1025.

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

Hedrick, M., Ohde, R., 1993. Effect of relative amplitude of frication on perception of place of articulation. J. Acoust. Soc. Amer. 94, 2005–2026.

Hermansky, H., 1990. Perceptual linear predictive (plp) analysis of speech. J. Acoust. Soc. Amer. 87, 1738–1752.

Hermansky, H., 1998. Should recognizers have ears? Speech Comm. 25, 3–27.

Hermansky, H., Morgan, N., 1994. Rasta processing of speech. IEEE Trans. Speech Audio Process. 2, 578–589.

Houtgast, T., Steeneken, H., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Amer. 67, 318–326.

Jongman, A., 1989. Duration of frication noise required for identification of english fricatives. J. Acoust. Soc. Amer. 85, 1718–1725.

Kewley-Port, D., Pisoni, D., Studdert-Kennedy, M., 1983. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. J. Acoust. Soc. Amer. 73, 1779–1793.

Kryter, K., 1962a. Methods for the calculation and use of the articulation index. J. Acoust. Soc. Amer. 34, 1689–1697.

Kryter, K., 1962b. Validation of the articulation index. J. Acoust. Soc. Amer. 34, 1698–1702.

Lee, B., Hasegawa-Johnson, M. (2007). Minimum mean-squared error a posteriori estimation of high variance vehicular noise. In: Biennial on DSP for In-Vehicle and Mobile Systems.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512.

Müsch, H., 2000. Review and computer implementation of Fletcher and Galt's method of calculating the articulation index. Acoust. Res. Let. Online 2, 25–30.

Omar, M.K., Hasegawa-Johnson, M., 2004. Model enforcement: A feature transformation framework for classification and recognition. IEEE Trans. Signal Process. 52, 2701–2710.

Padmanabhan, M., Dharanipragada, S., 2005. Maximizing information content in feature extraction. IEEE Trans. Speech Audio Process. 13, 512–519.

Pavlovic, C.V., Studebaker, G.A., 1984. An evaluation of some assumptions underlying the articulation index. J. Acoust. Soc. Amer. 75, 1606–1612.

Phatak, S., Allen, J., 2007. Consonant and vowel confusions in speech-weighted noise. J. Acoust. Soc. Amer. 121, 2312–2326.

Phatak, S., Lovitt, A., Allen, J., 2008. Consonant confusions in white noise: Effects of noise spectrum and utterance variability. J. Acoust. Soc. Am. 124, 1220–1233.

Remez, R., Rubin, P., Pisoni, D., Carrell, T., 1981. Speech perception without traditional speech cues. Science 212, 947–950.

Repp, B., 1986. Perception of the [m]-[n] distinction in cv syllables. J. Acoust. Soc. Amer. 79, 1987–1999.

Repp, B., 1988. Perception of the [m]-[n] distinction in vc syllables. J. Acoust. Soc. Amer. 83, 237–247.

Ronan, D., Dix, A., Shah, P., Braida, L., 2004. Integration across frequency bands for consonant identification. J. Acoust. Soc. Amer. 116, 1749–1762.

Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., Ekclid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Sharf, D., Hemeyer, T., 1972. Identification of place of consonant articulation from vowel formant transitions. J. Acoust. Soc. Amer. 51, 652–658.

Stevens, K., Blumstein, S., 1978. Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Amer. 64, 1358–1367.

Strope, B., Alwan, A., 1997. A model of dynamic auditory perception and its application to robust word recognition. IEEE Trans. Speech Audio Process. 5, 451–464.

Viemeister, N., Wakefield, G., 1991. integration and multiple looks. J. Acoust. Soc. Amer. 90, 858–865.