Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Speech Communication 45 (2005) 401–423

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Human and machine consonant recognition ☆

Jason J. Sroka [1], Louis D. Braida *

Research Laboratory of Electronics and Harvard-M.I.T. Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 9 February 2004; received in revised form 12 November 2004; accepted 22 November 2004

Abstract

Three traditional ASR parameterizations matched with Hidden Markov Models (HMMs) are compared to humans for speaker-dependent consonant recognition using nonsense syllables degraded by highpass filtering, lowpass filtering, or additive noise. Confusion matrices were determined by recognizing the syllables using different ASR front ends, including Mel-Filter Bank (MFB) energies, Mel-Filtered Cepstral Coefficients (MFCCs), and the Ensemble Interval Histogram (EIH). In general the MFB recognition accuracy was slightly higher than the MFCC, which was higher than the EIH. For syllables degraded by lowpass and highpass filtering, automated systems trained on the degraded condition recognized the consonants as well as humans. For syllables degraded by additive speech-shaped noise, none of the automated systems recognized consonants as well as humans. The greatest advantage displayed by humans was in determining the correct voiced/unvoiced classification of consonants in noise.

© 2005 Elsevier B.V. All rights reserved.

PACS: 43.71.E; 43.71.G; 43.72.D; 43.72.N

Keywords: Automatic speech recognition; Consonant identification; Filtering; Noise; Speech recognition

1. Introduction

Despite significant advances in Automated Speech Recognition (ASR) systems, performance at human levels has not yet been attained. Human recognition results provide proof that continuous speech can be recognized more accurately than the best current ASR systems. In theory, if a complete model of human speech processing were available, human-level performance would be

☆ Please do not reference or quote without permission. Supported by NIH Grants R01 DC00117 and T32 DC00038.

* Corresponding author. Address: 36-747 M.I.T, Cambridge, MA 02139, USA. Tel.: +1 617 253 2575; fax: +1 617 258 7003.

E-mail addresses: jjsroka@email.msn.com (J.J. Sroka), ldbraida@mit.edu (L.D. Braida).

[1] Current address: Alphatech Inc., Burlington, MA 01803, USA.

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.11.009

immediately realizable. While no such model is currently available, knowledge of how humans go about the speech recognition task is potentially useful in directing research on improving ASR systems.

Lippmann (1997) compared results of human and machine speech recognition and found that machine word error rates were typically about an order of magnitude greater in quiet environments. The gap in performance between humans and machines tends to get larger for speech in noise or when the recognition task gets more complex, for example, as the task moves from recognition of isolated words to continuous speech. Two sources of the superior recognition performance are better abilities for recognizing the features in the speech signal that carry information differentiating between phones and the use of higher-level speech mechanisms (e.g. vocabulary, syntax) for delivering information.

This paper continues a line of research comparing human and machine speech recognition performance at the consonant recognition level in tests that minimize or remove higher-level (e.g. lexical, syntactic) language mechanisms for information flow between a speaker and a listener. By removing these effects, the comparison explores how well humans and machines can determine the speech signal features that distinguish between a set of consonants. Comparisons are made in a range of degraded conditions enabling examination of recognition as the audio cues are increasingly masked or removed. Examination of common confusions explores which differentiating features are most robustly identified by the humans and the automated systems. By testing a number of front ends we can examine if there are differences in their ability to provide differentiating features to a common back end and whether incorporating knowledge of human audio processing leads to more human-like error patterns.

Much work in this area has focused on using models of auditory nerve encoding of speech as the front end or parameterization stage of ASR. One such auditory-based front end is the Ensemble Interval Histogram (EIH) as described in Ghitza (1994). Ghitza (1993) compared performance of cepstral coefficients calculated using the Fourier Power Spectrum with cepstral coefficients calculated using the auditory-based Ensemble Interval Histogram (EIH) front end and with human performance. The speech task was the Diagnostic Rhyme Test (DRT). The DRT is structured to measure relative performance on recognition of phonetic features, which differentiate between consonant sounds. Recognition was tested for speech in additive flat-spectrum noise at Speech to Noise Ratios (SNRs) of +30, +20, and +10 dB.

While the cepstral coefficients based on the EIH outperformed those based on the Fourier Power Spectrum in noisy conditions, neither automated system approached human-level performance. The disparity in overall recognition levels and the lack of errors by human listeners for half of the error categories examined makes response pattern comparisons problematic.

Jankowski et al. (1995) compared performance of a traditional mel-cepstra system with Seneff's synchrony/mean-rate model and with Ghitza's EIH in additive speech babble over a range of Speech to Noise Ratios (SNRs). The speech task was isolated word recognition for a set of 105 aircraft commands. Results from the systems are generally similar except at the poorest SNR (+6 dB), where the error rate for the auditory-based systems was roughly 22% compared to 27% for the traditional mel-cepstra system.

Human performance levels for recognition of noise-free, read speech (the CSR'94 Spoke 10 corpus, made up of multiple speakers reading passages from the 5000 word vocabulary Wall Street Journal corpus) do not vary as noise is added down to +10 dB SNR (Ebel and Picone, 1995). Jankowski et al. found that error rates of the automated systems they tested increase from roughly 1.0% in quiet to 7.5% at +12 dB SNR with training on clean speech alone. When training included speech degraded by noise, error rates increased from less than 1% in quiet to roughly 3.5% at +12 dB SNR. This is an example of the sensitivity of machine performance to mismatches between training and test conditions.

This paper compares the performance of the auditory-based EIH front end and two traditional front ends, Mel-Filter Bank (MFB) energies and Mel-Frequency Cepstral Coefficients (MFCCs) with data on human performance on the same task: recognition of consonants in nonsense syllables. This extends comparisons between humans and machines at the subword recognition level over degradations sufficient to cause a significant number of errors for the human listeners and thereby allows for comparisons of error patterns between humans and machines.

The paper is organized as follows. The databases and outcome measures used, and the human consonant perception tests, are described in Section 2. The implementation of the ASR systems is described in Section 3. Comparisons, described in Section 4, consider both recognition scores and error patterns to determine which front ends, if any, produce error patterns similar to those of humans. Finally, Section 5 contains a summary and discussion of the results.

2. Methods

2.1. Speech materials

Two sets of human consonant recognition results were used with one set of machine consonant recognition results for comparing human and machine recognition. One set of human results is for a database composed of consonant–vowel–consonant (CVC) syllables and used to test human recognition in highpass and additive noise conditions. This same CVC database is also used for all machine recognition testing (highpass, lowpass, and additive noise). The other set of human results is reported by Miller and Nicely (1955) for consonant–vowel (CV) syllables, on which humans were tested in highpass, lowpass, and additive noise conditions. Because human results for recognition of lowpass filtered speech are not available for the CVC database, lowpass comparisons are limited to machine recognition on the CVC database with human results on Miller and Nicely's CV database.

2.1.1. CVC database

The CVC database is composed of 496 Consonant–Vowel–Consonant syllables, each preceded by a schwa. The materials were recorded by one male and one female talker, each of whom produced half of the CVC tokens, in the form /ə/-CVC, where /ə/ is the unstressed schwa. The syllables were constructed using 12 consonants and 6 vowels. The consonants were /p, t, k, b, d, g, θ, v, ð, s, ʃ, z/ and the vowels were /i, ɑ, u, ɪ, ɛ, ʊ/. The mean durations of the tokens spoken by each of the two talkers were 634 (M) and 574 (F) ms.

The initial and final consonants for a CVC token were independently drawn with probability 1/12 from the set of 12 consonants, allowing for duplications of CVCs (one male and one female token) and omissions. The first 12 rows of Table 1 lists the consonants that are present in the database, along with their classification with respect to four different distinctive features. The CVC database had been used in psychophysical tests of human phone recognition for speech processed by amplitude compression (Lippmann, et al., 1981; Bustamante and Braida, 1986; DeGennaro et al., 1986; Bustamante and Braida, 1987) and more recently for speech degraded by filtering (Ronan et al., 2004), and by additive speech-shaped noise (Dix and Braida, 2002). During human testing, listeners knew the constraints of the database (e.g. that whatever token was chosen from the CVC database, it would begin with a schwa and be followed by a Consonant–Vowel–Consonant sequence, where each of the consonants would be one of 12 possible, and each of the vowels would be one of six possible).

The speech was lowpass filtered to a bandwidth of 4500 Hz, sampled at 10 kHz, converted to 12 bit samples, and normalized to have the same rms value. In the tests on humans, filtering was performed by linear-phase FIR filters designed using Matlab (Mathworks, 2004). Filter lengths varied from 896 to 1216 points. Transition regions widths were 50 Hz, with out-of-band attenuations of at least 80 dB. Filter bands were 700–4500, 1400–4500, 2100–4500, and 2800–4500 Hz. Unfiltered speech was presented at 75 dB SPL. After filtering, speech-shaped noise (at +35 dB SNR) was added to the filtered signal to obscure the speech in the stop band.

In separate tests, speech-shaped noise was added to the unfiltered (0–4500 Hz, presented at 75 dB SPL) speech at SNRs of −7, −4, +2, +8,

Table 1
Feature descriptions of the 12 consonants used in the CVC database (first 12 lines) and of the additional four used by Miller and Nicely (1955, final four lines)

Consonant	Voicing	Frication	Sibilance	Place	Example
p	0	0	0	0	peep
t	0	0	0	1	toot
k	0	0	0	2	kick
b	1	0	0	0	bob
d	1	0	0	1	deed
g	1	0	0	2	gig
θ	0	1	0	1	thin
v	1	1	0	0	vet
ð	1	1	0	1	them
s	0	1	1	1	sis
ʃ	0	1	1	2	shin
z	1	1	1	1	zip
f	0	1	1	0	fat
ʒ	1	1	1	2	azure
m	1	0	0	0	mom
n	1	0	0	1	now

For voicing, frication, and sibilance, a 0 indicates the feature is absent and a 1 indicates the feature is present. The place classes follow those of Miller and Nicely: 0 = labial, 1 = alveolar, and 2 = velar. Note that the Sibilance feature here corresponds to Miller and Nicely's 'Duration' feature.

and +14 dB. Tests by five listeners who did not participate in the experiments reported in this paper indicated that initial consonants were recognized correctly 90.9% of the time and final consonants 89.4% of the time for speech that was not degraded by noise.

For the ASR tests using highpass filtering, somewhat different filters were used. All filters used 512 coefficients to create sharp cutoffs equivalent to roughly a sixth-order highpass or lowpass filter at the cutoff frequencies. Highpass filtering was applied using cutoffs of 200, 400, 800, 1600, and 2800 Hz. Lowpass filtering was applied using cutoffs of 500, 1000, 2000, 2400, 2700, 3000, and 4000 Hz. After filtering, speech-shaped noise (at +30 dB SNR) was added to the filtered signal to obscure the speech in the stop band.

For ASR tests of speech degraded by additive speech-shaped noise, a Grason–Stadler noise generator (model 901B) was used to generate noise with a spectrum that was flat to 1 kHz and then fell at roughly 6 dB/octave. This was the same generator and same spectrum as was used in the psychophysical tests of CVCs. Noise was added to the CVC tokens to create databases at SNRs of 0, +5, +10, +20, and +30 dB. To generate a signal at a desired SNR, the root-mean-squared (rms) values of the normalized CVC tokens (0–4500 Hz bandwidth) were calculated over entire CVC database and then noise was added at an appropriate rms level relative (0–4500 Hz bandwidth) to the overall average rms value. This is consistent with the method used for the human CVC studies.

2.1.2. CV database

The CV database (Miller and Nicely, 1955) was composed of 16 consonants, including /f/, /ʒ/, /m/, and /n/ in addition to all 12 present in the CVC database (see Table 1), but only the vowel /a/ as in "father" in contrast to the six vowels present in the CVC database. Tests explored recognition in highpass filtering conditions (actually bandpass conditions of 1000–, 2000–, 2500–, 3000–, 4500–5000 Hz) and in lowpass filtering conditions (actually bandpass conditions of 200–300, –400, –600, –1200, –2500, –5000, and –6500 Hz). The highpass and lowpass filter skirts were 24 dB per octave, and the filtered speech was further degraded by adding flat-spectrum noise at +12 dB SNR. By contrast, the CVC tests used much less noisy speech for the filtering tests. Miller and Nicely also performed tests of wideband (200–6500 Hz) speech in additive flat-spectrum noise (at SNR of -18, -12, -6, 0, +6, and +12 dB).[2]

Miller and Nicely (1955) tested five female subjects. One subject would speak randomly-ordered syllables and the other four would record their responses. The speaker would alternate with being a listener in the experiments. All of the Miller and Nicely subjects were United States citizens with the exception of one Canadian, and none had a noticeable dialect. Miller and Nicely thus did not use recorded syllables. Instead, speakers were trained to produce CV syllables at a "constant" amplitude (standard deviation of 1.04 dB). The average peak amplitude over a number of examples was then used as the basis for calculating necessary noise levels for the range of SNRs tested.

We report comparisons between the Miller and Nicely CV results with the human and machine CVC results for highpass and additive noise conditions. We shall show that the results from the Miller and Nicely CV study are similar in many ways to our human CVC results. Because no studies of human recognition of the CVC database in lowpass conditions exist, lowpass comparisons will be limited to machine results on the CVCs with Miller and Nicely's CV results.

2.2. Outcome measures

Performance of the human and machine systems was evaluated in terms of percentage of correct responses and also in terms of information transfer.

A confusion matrix was created from the responses made by all four listeners to both initial and final consonants, ignoring the response to the vowel. The Identificaton Score (percentage of correct responses) was computed from the entries in this confusion matrix, $\{N_{ij}\}$.

$$P = 100\frac{\sum_{i=1}^{M} N_{ii}}{\sum_{i=1}^{M}\sum_{j=1}^{M} N_{ij}} \tag{1}$$

where M is the number of consonant items.

An information transfer measure was also computed for certain subdivisions of the confusion matrix. The Relative Information Transfer score is computed by mapping the confusion matrix into an $F \times F$ feature matrix, $\{K_{ij}\}$ and then computing H and I, defined as

$$H = \sum_{i=1}^{F} \frac{K_{i\bullet}}{K_{\bullet\bullet}} \log\frac{K_{\bullet\bullet}}{K_{i\bullet}} \tag{2}$$

$$I = 100\frac{1}{H}\sum_{i=1}^{F}\frac{K_{i\bullet}}{K_{\bullet\bullet}}\sum_{j=1}^{F}\frac{K_{ij}}{K_{i\bullet}}\log\left(\frac{K_{ij}}{K_{i\bullet}}\right)\left(\frac{K_{\bullet\bullet}}{K_{\bullet j}}\right) \tag{3}$$

where F is the number of feature categories, H is the entropy in the stimulus set, I is the Relative Information Transfer and[3]

$$K_{i\bullet} = \sum_{j=1}^{F} K_{ij} \tag{4}$$

$$K_{\bullet j} = \sum_{i=1}^{F} K_{ij} \tag{5}$$

$$K_{\bullet\bullet} = \sum_{i=1}^{F}\sum_{j=1}^{F} K_{ij} \tag{6}$$

The quantities H and I are both positive, with

$$H \leqslant \log N \tag{7}$$

and

$$0 \leqslant I \leqslant 100 \tag{8}$$

In general, information transfer provides a measure of the consistency of responses conditioned on the presentation of a stimulus rather than their correctness. More specifically, it measures the extent to which a stimulus can be uniquely identified given knowledge of the response. Information transfer is not simply a monotonic transform of the percentage of correct responses: the two differ in the way in which different types of responses are weighted.

[2] The difference between the CVC bandwidth, 0–4500 Hz, and the CV bandwidth, 200–6400 Hz, corresponded to about a 12% reduction in the Articulation Index (ANSI, 1997) for noiseless speech. The difference is somewhat less than this because Miller and Nicely tested at +12 dB SNR.

[3] The terms in these expressions, e.g. $\frac{K_{ij}}{K_{i\bullet}}$ and $\frac{K_{\bullet\bullet}}{K_{\bullet j}}$, correspond to the terms in the confusion matrices presented in Tables 3–11.

While one could compute information transfer scores for the entire confusion matrix (such a computation provides a useful alternative to the overall percent correct score), it is far more interesting to divide a matrix into parts by phonetic features and compute the information transfer score for the different parts. For example, a 2 × 2 confusion matrix can be constructed for Voicing by segregating stimuli and responses according to whether they are Voiced or not. In this case, errors only correspond to the confusion of an Unvoiced consonant for a Voiced one and *vice versa*. A Relative Information Transfer measure is used (Miller and Nicely, 1955; Bratakos et al., 2001) to permit different numbers of categories to be compared, unlike mutual information.

The features used for analyzing the consonant set were the same as those used by Miller and Nicely except we omit Nasality since none of the CVC consonants were Nasals. The four features used were Voicing, Frication, Sibilance (corresponding to Miller and Nicely's "Duration"), and Place. Using these features allows each of the CVC consonants to have a distinct combination of feature values. The consonants are evenly divided between Voiced and Unvoiced (six and six as shown in Table 12). Half of the consonants are Fricatives,[4] with a subset of three Sibilant Fricatives. When the consonant set is divided by the three Place values, there are three labials, six alveolars, and three velars.

It should be noted that, as pointed out by Miller and Nicely, the features are generally not independent. For example, in the set of 12 consonants that we tested, all Sibilant consonants were Fricatives. An exception to this statement applies to the Voicing and Frication features in the case of the CVC syllables. For all other pairs of features, knowledge of the value of one feature provides information about the value of another feature for the CVC dataset.

2.3. Human consonant perception tests

Two groups of listeners were tested on the CVC database, one on filtered speech and one on speech degraded by noise. Both groups were tested without correct answer feedback, after a period of training on non-test items in which feedback was presented. The first group consisted of 2 M and 3 F, 18–22 years old, the second consisted of 1 M and 2 F, 19–22 years old. For all subjects, English was their primary language. Results were combined across vowels because it was desired to obtain measures of identification that were broadly representative of vowel context.

Percent correct scores of both groups were subject to Anova analysis (Winer, 1971). For the group tested on filtered speech, although Consonant Position (initial or final) Filter, and Subject all had significant effects (at the 0.01 level of significance), Filter × Subject, Filter × Position, and Position × Subject did not. Consequently, results were averaged across subject and position to obtain the matrices to analyze. For the group tested on speech degraded by noise, although Position, Speech to Noise Ratio (SNR), and Subject had significant effects (at the 0.01 level) Subject × Position and Subject × SNR did not. While the effect Position × SNR was significant, it accounted for only 1.1% of the variance. Consequently, results were averaged across both subject and position.

3. The automated recognition systems

A Hidden Markov Model back end was paired with three front ends: the Mel-Filter Bank (MFB) energies, the Mel-Filter Cepstral Coefficients (MFCCs), and the Ensemble Interval Histogram (EIH). The next sections describe the MFB and MFCC parameterizations (Section 3.1), the EIH parameterization (Section 3.2), and the HMM back end (Section 3.3).

3.1. The MFB and MFCC parameterizations

The MFB and MFCC front ends were implemented using the Hidden Markov Model Toolkit software package (HTK, 2004). Each of these

[4] In the CVC database, but not in the CV database, all the Non-Fricative sounds are Plosives and all the Fricatives, Non-Plosives, so all reports of Frication scores for the CVC database could be equally applied to the feature "Plosive".

front ends generated a set of parameters every ten milliseconds using a twenty millisecond Hamming window for analysis of the input signal. The ten millisecond frame rate and twenty millisecond analysis window values are within the ranges typically used for ASR systems (Jankowski, 1995).

The Mel-Filter Bank (MFB) front end was implemented by binning the squares of the magnitudes the Fourier transform coefficients of the speech. The coefficients were binned by multiplying by the filter gain and the results accumulated. Thus, each bin holds a weighted sum representing the spectral power in that filterbank channel. The basic filter shape is triangular, with unit magnitude at its center and extending to the center (peak) of the neighboring filters. The filters were equally spaced along the a mel-warped frequency axis:

$$\mathrm{Mel}(f) = 2595\log_{10}(1 + f/700) \tag{9}$$

where f is frequency in Hz.

The 24 spectral energies were supplemented by 24 Δ-coefficients, weighted measures of the rate of change for each spectral energy parameter over a 50 ms time span. In addition, an overall energy term was calculated. This energy value was only used to calculate a Δ value (corresponding to change in overall energy) and was not used directly as a parameter. This yielded a final 49-element parameter vector (24 MFB values, 24 Δ-MFB values, and 1 Δ-energy value (Rabiner and Juang, 1993).

The MFB can be seen as a fairly straightforward model of auditory processing. It implements a spectral energy analysis along a frequency axis warped to imitate the frequency mapping along the basilar membrane and supplements the representation with information about the pattern of changes over 50 ms centered on the analysis time.

The MFCC front end used a Discrete Fourier Transform to convert the mel-warped spectra into 24 cepstral coefficients for each frame of speech. The zeroth cepstral coefficient (corresponding to the signal energy in the frame) was used in the same way as the energy value in the MFB system. This resulted in a final 49-element parameter vector (24 cepstral coefficients, 24 Δ-cepstra, and 1 Δ-energy value).

3.2. Ensemble Interval Histogram (EIH)

The EIH (Ghitza, 1994) uses a set of bandpass filters developed to display tuning characteristics similar to neurons in the auditory nerve. After speech is bandpass filtered, threshold detectors produce a neural event every time a positive crossing of a threshold occurs. Four (or five) thresholds cover the 30 dB range of normal speech. After converting the speech into a set of threshold crossing times on the various channels, the intervals between consecutive crossings are measured. A histogram of frequencies corresponding to the reciprocals of the intervals is then collected across thresholds and across channels. This histogram is termed the Ensemble Interval Histogram (EIH).

Implementation of the EIH front end can be divided into the stages of waveform filtering, determination of individual threshold crossings in model channels, and calculation of the intervals between threshold crossings. The filters used in the implementation were generated with a Matlab toolkit (Slaney, 1998). The output of the filtering stage resulted in a 73-channel representation of the speech input. Each channel was then half-wave rectified. The next stage determined the times that a positive threshold crossing occurred for each of seven logarithmically-spaced thresholds for each channel. The original EIH representation varied the thresholds for each channel of the filtering output to match amplitude ranges in individual channels. In this research a single set of thresholds was used for all channels in order to reduce computational complexity. The thresholds were logarithmically spaced to cover the range of amplitudes exhibited by low-frequency channels for some sample CVCs. This resulted in a 511 bit per-frame representation (73 channels × 7 thresholds per channel) where each sample had a value of 1 (when the threshold for that channel had been crossed positively) or zero (when the threshold had not been crossed positively).

Intervals between consecutive threshold crossings in individual channels were then recorded. To determine the EIH representation for a particular frame, each channel was examined over a length of time equal to 10 times the reciprocal of the channel's Center Frequency (CF, calculated as the frequency corresponding to the peak in the frequency response curve for the filter used to generate the channel) back in time with a maximum window length of 40 ms.

A histogram with 24 bins (chosen to be equivalent to the number of filters used in the MFB representation) was constructed based on the number of samples in the intervals between threshold crossings. The bins were determined by logarithmically spacing frequency thresholds between roughly 100 and 5000 Hz with the following constraints added to the procedure. In order not to increase the computational load by requiring upsampling or interpolation, bins were rounded to an integer number of samples (the waveforms were sampled at 10 KHz, so samples were spaced 0.1 ms apart). An obvious additional constraint was that no two bins round to the same number. Table 2 shows the sample intervals per bin. These constraints were not used by Ghitza in his original EIH implementation. Their effect is to give more weight to the low frequencies than either Ghitza's original EIH or either of the two traditional parameterizations (MFB or MFCC). The 24 EIH Δ-parameters and 1 Δ-energy value were calculated identically to the MFB and MFCC parameterization.

Table 2
Neural event intervals for each EIH bin

Bin #	Samples
1	85–97
2	73–84
3	63–72
4	55–62
5	47–54
6	41–46
7	35–40
8	30–34
9	26–29
10	23–25
11	20–22
12	17–19
13	15–16
14	13–14
15	11–12
16	10
17	9
18	8
19	7
20	6
21	5
22	4
23	3
24	2

3.3. The hidden Markov model back end

The same Hidden Markov Model (HMM) back end (Rabiner and Juang, 1993) was paired with each of the parameterizations (MFB, MFCC, and EIH) discussed above. For all the results reported here, the HMMs had five states per phone model with fifteen diagonal covariance Gaussians per phone model state. Transition matrices allowed skipping states within a phone model.

The performance of ASR systems, especially in degraded signal conditions, is dependent on the speech used to train the systems. Systems trained and tested under unmatched conditions have usually displayed poor performance. To minimize these effects, we used Degradation-Specific Training (DST): a system is trained and tested on the same degradation (e.g. additive noise at 5 dB SNR).

We also evaluated the systems with Clean Training (CT) in which the systems were trained on full bandwidth noiseless speech, and Mixed Training (MT), in which the systems were trained on a mixture of the filtering conditions or a mixture of the additive noise conditions. In general performance with Mixed Training was slightly below DST levels, while Clean Training results were much worse than either DST or MT. With one exception, the orderings of performance for the various systems were the same as for the DST training. The exception was for CT with lowpass filtering, for which the performance of the EIH system exceeded that of the MFB and MFCC systems for lowpass filtering more severe then 2400 Hz.

For training purposes, the CVC database was split into three token sets of equivalent sizes, labeled DB1, DB2, and DB3. The token sets were selected such that both initial and final consonants were evenly distributed across the three sets. ASR systems were trained on two of the three sets and tested on the third (e.g. trained on DB1 and

DB2, and tested on DB3) using each of the three sets as a test set once.

With the DST approach, a separate system needs to be trained for each type and severity of degradation (e.g. highpass filtering with a 200 Hz cutoff frequency), with separate systems for each of the three subdivisions of the database. Thus, for the 200 Hz highpass filtering condition, three MFB systems were trained and tested, one on each of the three divisions of the speech data (DB1, DB2, and DB3 as described above), using the same number of states and mixtures per state for each MFB system but allowing training to determine the particular parameter values. The results reported reflect the average of the three separate training and testing sets for a particular degraded condition.

Human training is difficult to match to any common ASR training regimen. Nevertheless, humans perform very well even in relatively "untrained" conditions (i.e. conditions which are not commonly experienced such as highpass filtering at 2 kHz). This demonstrates an ability to handle or adapt to novel environments, which machines have not shown.

During the training and recognition phases, the machine systems used a phone grammar reflecting the sequence silence–vowel–consonant–vowel–consonant–silence. All sequences of phones matching this pattern were equally likely. This approach removed the possibility of insertions and deletions as error types and removed a significant source of variability from the general speech recognition task.

By comparing human and machine performance, we hope to determine if these parameterizations matched with HMMs are able to extract cues in the degraded speech signals that allow human-level recognition for phones or phonetic features. Significant advantages are provided to the machine systems, including the limited speech material (creating effectively speaker-dependent systems) and the phonemic constraints (knowing in advance the phone sequence was vowel–consonant–vowel–consonant), so that the machine performance levels are likely near the best performance levels possible on this speech data. The human CVC results presented herein are for subjects who had analogous advantages.

4. Performance comparisons

In this section, human performance on the CVC and CV databases is compared to machine performance on the CVC database for highpass filtering, lowpass filtering, and additive noise degradations. Identification scores are first compared. Next, Relative Information Transfer scores are reported for divisions of the consonant set based on phonetic features.

4.1. Confusion matrices

Normalized confusion matrices are presented in Tables 3–11 for the human CVC results and in Tables 12 and 13 for the machine CVC results. In these confusion matrices, the consonants are ordered so that the Plosives come before the Fricatives. Relative Information Transmission scores for Frication, then, view a confusion matrix as four equal-sized (6 by 6) quadrants, with the upper left quadrant (Plosive presented, Plosive recognized) and the lower right quadrant (Fricative presented, Fricative recognized) considered correct and the upper right (Plosive presented, Fricative recognized) and lower right (Fricative presented, Plosive recognized) considered incorrect. The Sibilant subset of the Fricatives are grouped within the Fricative set such that they are the final three consonants. The Relative Information Transfer score for Sibilance will again divide the confusion matrix into quadrants, but for Sibilance the quadrants will not be of equal size: the upper left quadrant will be 9 by 9 and the lower right 3 by 3.

Two machine recognition confusion matrices are included for comparison: one in which the degradation was 1600 Hz highpass filtering (Table 12) and the other in which the degradation was additive noise at +10 dB SNR (Table 13). These may usefully be compared with confusion matrices from humans for highpass filtering at 1400 Hz (Table 4) and in additive noise at +2 dB SNR (Table 9). These comparisons are of interest because they allow a range of Relative Information Transfer scores to be examined in matrices with comparable consonant error rates (ranging from 70% to 80%).

The MFB identification score was 80% for the 1600 Hz highpass filtered speech, which yielded

Table 3
Human results on the identification of 700 Hz highpass-filtered CVC syllables

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	94.5	0.8		3.0		0.8					0.8		363
t	0.9	97.7	0.3	0.6					0.6				343
k		0.9	42.3				0.3	0.3	0.6	25.7	27.7	2.3	350
b	0.3	0.9		96.8	2.1								339
d				0.3	98.0	0.3			1.4				358
g	1.7				2.6	95.4				0.3			349
θ		0.3	0.3				82.6	6.4		10.4			357
v							5.2	93.9		0.9			344
ð		0.3			1.1				98.6				352
s			0.3				6.4		0.3	91.5	1.2	0.3	343
ʃ	0.3	0.3	16.1				0.3			3.2	79.5	0.3	342
z			2.3				2.9	0.3		1.1	4.3	89.1	350
all	8.4	8.3	5.1	8.2	8.9	8.0	8.3	8.3	8.5	11.0	9.3	7.7	4190

The first 12 rows each correspond to a given transmitted phone and the entries in the various columns correspond to the percentage of times that the phone was responded when the transmitted phone was uttered. Blank entries indicate that percentage of times was less than 0.2%. The thirteenth column indicates the total number of times each stimulus was presented. The thirteenth row corresponds to average over all transmitted phones.

Table 4
Human results on 1400 Hz highpass-filtered CVC syllables

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	78.1	4.1		9.6		3.5			0.3	0.6	3.5	0.3	342
t	0.3	94.4		1.4					3.9				360
k	1.0	1.3	37.0	1.3			1.0			28.2	27.6	2.6	308
b	6.1	0.9		89.3	3.4						0.3		326
d	0.3		0.6	4.1	89.0	4.1			1.9				363
g	4.8	1.5		3.3	13.7	75.0			1.5		0.3		336
θ	0.3						68.6	12.8		16.6	0.3	1.5	344
v		0.3					5.8	90.2		2.4	0.3	0.9	328
ð		4.1		0.3	2.9	1.2			91.3	0.3			344
s			2.7				5.1		0.9	87.2	3.3	0.9	335
ʃ	2.1	0.9	17.0	0.6						14.0	62.9	2.4	329
z			1.8				4.2	0.6		1.2	5.7	86.4	331
all	7.8	9.5	4.6	8.9	9.6	7.0	7.1	8.5	8.5	12.3	8.4	7.8	4046

See Table 3 for details.

the highest Relative Information Transfer score of 98% when recognizing Sibilance. Dividing the matrix (Table 12) into Sibilance-based quadrants as described above, we see the error rate was very low, (/θ/ recognized incorrectly as /s/ 0.7% of the time, and /θ/ recognized incorrectly as /z/ 2.1% of the time). By comparison, the results for humans on a slightly less severe degradation (Table 4) were an identification score of 72% and a Sibilance score of 66%, consistent with the much more numerous errors.

The MFB system recognizing speech at a SNR at +10 dB SNR (Table 13) had an identification score of 74% but a Relative Information Transfer score for Sibilance of 96%. The drop in Sibilance score from 98% for 1600 Hz highpass

Table 5
Human results on 2100 Hz highpass-filtered CVC syllables

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	47.5	16.8	1.1	21.3	3.5	2.7			0.5	1.9	4.8		375
t	0.9	83.6	0.6	9.2	0.9		0.3		4.5				336
k	0.9	3.6	34.1	2.1				0.3	0.3	18.1	33.5	7.1	337
b	8.5	7.3	0.6	76.9	3.2	0.6			1.2	0.9	0.9		342
d	0.9	0.9		6.9	70.3	11.0			9.5	0.3	0.3		347
g	8.0	3.5		9.7	24.8	45.4			6.8	1.5	0.3		339
θ							30.5	60.4		2.3		6.7	341
v			0.3				9.1	83.4		3.3		3.9	362
ð		2.0		0.3	4.6				92.8	0.3			348
s	0.3	0.3	3.0		0.6	0.3	4.5	4.2	1.2	77.5	6.9	1.2	334
ʃ	1.5	1.8	23.4	1.8			0.3	0.3	0.3	10.8	55.0	5.0	342
z		0.6	4.1	0.3			4.1	3.8		1.7	5.5	79.9	343
all	6.0	9.9	5.5	10.8	9.0	4.9	4.1	13.0	9.8	9.6	8.8	8.6	4146

See Table 3 for details.

Table 6
Human results on 2800 Hz highpass-filtered CVC syllables

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	47.0	18.4	3.2	7.1	0.8	9.5	0.2	0.2	2.1	4.0	7.2	0.2	956
t	5.1	71.0	0.1	11.8	0.6	2.6	0.1	0.1	7.5	0.8	0.3		968
k	2.6	3.3	23.5	0.8		0.3	0.5		0.7	17.6	46.3	4.4	962
b	16.4	28.3	1.1	36.4	3.3	3.3	0.1		3.9	2.8	4.3		966
d	2.8	4.6	0.3	3.1	34.3	18.7	0.2	0.2	30.3	3.1	1.8	0.6	967
g	6.5	4.9	0.4	4.4	11.9	52.1		0.2	14.7	2.7	2.0	0.1	971
θ	0.1	0.2	0.1	0.1			35.3	57.5	0.2	1.9	0.1	4.5	985
v	0.1	0.2	0.3		0.1		11.9	81.7	0.3	1.8	0.2	3.4	977
ð	0.3	5.2	0.1	0.6	3.1	2.0	0.1	0.2	87.7	0.3		0.4	965
s	0.7	0.7	5.3	0.6	0.2	0.4	7.3	4.7	0.9	67.9	7.6	3.6	951
ʃ	1.7	2.7	19.4	1.8	0.2	0.5	0.9	0.4	0.6	17.3	48.9	5.6	960
z	0.1		2.4	0.4	0.2	0.1	5.9	8.2	0.2	4.1	4.3	74.1	968
all	6.9	11.6	4.7	5.6	4.6	7.5	5.3	12.9	12.4	10.3	10.2	8.1	11,596

See Table 3 for details.

filtering (Table 13) to 96% reflects slightly more Sibilance errors in the noisy condition. By comparison, the human results when recognizing noisy speech at a SNR at +2 dB SNR (Table 9) indicate an identification score of 73% and a Sibilance score of 87%, again reflecting the larger number of human Sibilance errors.

For Frication (or Plosiveness) essentially the same pattern of results occurs. The Relative Information Transfer (percentage correct) scores were 88% (98%) for the MFB system for 1600 Hz highpass filtering and 58% (91%) for humans for 1400 Hz highpass filtering; scores were 61% (92%) for the MFB system at a SNR of +10 dB and 47% (88%) for humans at a SNR of +2 dB.

4.2. Identification scores

Fig. 1 compares human identification scores obtained with the CVC and CV databases on

Table 7
Human results for CVC syllables degraded by noise at S/N = +14 dB

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	90.9	2.0	5.1	1.0			1.0						197
t		99.0			0.5		0.5						200
k		1.0	99.0										204
b	1.5		0.5	87.4	2.0	1.5	0.5			6.1	0.5		198
d		0.5		0.5	97.7	0.5		0.5		0.5			217
g					1.6	95.8	0.5			1.0	1.0		191
θ		3.5		1.2			83.2	0.6	0.6	1.2	9.8		173
v							6.9	83.1	10.1				189
ð								4.2	95.3		0.5		192
s	0.5	0.5	0.5	3.7			1.1			82.6	10.5	0.5	190
ʃ				0.6	1.1	1.7	13.8			38.5	43.7	0.6	174
z							0.6	2.2	1.1	6.7	1.1	88.3	179
all	7.9	9.2	9.3	8.1	9.6	8.2	8.2	7.4	8.9	11.0	5.2	6.9	2304

See Table 3 for details.

Table 8
Human results for CVC syllables degraded by noise at S/N = +8 dB

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	82.5	1.5	8.8				6.2				1.0		194
t		100.0											193
k	3.9	1.0	94.1				1.0						204
b	0.5			85.1	3.7	0.5	0.5			6.4	3.2		188
d		0.5		1.6	95.3	0.5				0.5	1.6		193
g				1.4	0.9	94.8	0.5			0.9	1.4		212
θ	1.2	4.2	1.2				84.9	1.2			7.2		166
v							6.1	83.3	10.6				198
ð							0.5	2.7	96.8				185
s				8.7	1.1	1.6	2.7			71.6	13.1	1.1	183
ʃ				1.1	4.4	2.2	8.8			41.4	41.4	0.6	181
z							0.5	2.0	0.5	4.4	1.0	91.7	205
all	7.4	8.9	9.2	8.0	8.8	9.1	8.3	7.6	8.7	10.0	5.5	8.3	2302

See Table 3 for details.

highpass, lowpass, and noisy speech with identification scores from the EIH, MFB and MFCC automated systems.

The upper left panel compares the two sets of human identification scores with the results from the three automated systems for the highpass filtering conditions. For the machine results, the MFB and MFCC identification scores are very similar. The EIH system does not do as well in these highpass conditions. This is likely due at least in part to the constraints placed on threshold crossing interval bins, which provided fewer separate bins for high frequency channels when compared to the MFB and MFCC systems.

Over the range they can be compared, the human CVC results are 10–20 percentage points better than the human CV results. In part this reflects the effects of the difference between the +35 dB SNR used for the CVC tests and the +12 dB SNR used by Miller and Nicely for their filtering studies. Thus their results are expected to be somewhat lower than the CVC results. The human CVC

Table 9
Human results for CVC syllables degraded by noise at $S/N = +2$ dB

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	65.1	9.5	14.2	1.0			9.4			0.4	0.6		727
t	0.4	98.3	1.0				0.3						763
k	9.2	6.3	80.7			0.3	3.1			0.5			782
b	0.5	0.2		54.4	6.3	4.5	2.6		0.1	25.5	5.4	0.5	820
d	0.1	0.8	0.1	1.9	83.2	4.7	1.2			5.0	2.9		782
g	0.4	0.3	0.1	7.8	5.9	70.8	1.1			10.2	3.3		791
θ	8.4	10.2	6.0	2.1	0.3		66.9	0.1		0.8	5.2		752
v							3.1	82.3	13.9	0.4	0.1	0.1	770
ð							2.5	2.8	94.4		0.3		785
s		0.4		12.5	5.7	4.7	2.4			64.5	9.5	0.3	786
ʃ		0.3		7.4	8.8	5.1	5.7		0.1	45.5	26.8	0.4	770
z			0.1		0.1	0.4		2.2	0.2	7.2	0.4	89.4	810
all	6.6	10.3	8.5	7.5	9.3	7.7	8.0	7.2	9.1	13.5	4.5	7.9	9338

See Table 3 for details.

Table 10
Human results for CVC syllables degraded by noise at $S/N = -4$ dB

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	43.5	14.0	23.3	1.6	0.5	0.5	14.0			1.6	1.0		193
t	2.1	94.2	2.1		0.5		1.0						191
k	19.2	10.5	54.8	0.9	0.5	0.5	11.9		0.9	0.9			219
b	2.3			27.1	11.3	8.5	7.3			34.5	9.0		177
d	0.6	0.6	0.6	6.1	56.4	7.7	5.5		0.6	13.3	7.7	1.1	181
g	1.7	0.6	0.6	9.2	13.9	31.8	7.5			26.6	7.5	0.6	173
θ	13.3	12.2	12.2	5.3	1.6	0.5	47.9	1.6	0.5	2.1	2.7		188
v							6.7	73.3	16.4	0.5	1.0	2.1	195
ð							8.2	3.5	88.3				171
s	0.5	0.5	0.5	16.7	7.4	8.3	2.8			52.3	10.2	0.9	216
ʃ	0.5	0.5		4.9	16.5	8.7	4.9	0.5		44.7	18.9		206
z					3.2	1.4	0.9	4.1	0.9	10.6	2.3	76.6	218
all	7.1	11.0	8.4	5.8	9.0	5.4	9.7	7.0	8.1	15.9	5.1	7.6	2328

See Table 3 for details.

results, taken at +35 dB SNR, are in agreement with the better machine results (MFB and MFCC), and considerably better than EIH. Both the MFB and MFCC results display human levels of performance, even surpassing human performance at the most severe highpass condition (2800 HZ cutoff). EIH results never exceed human levels on the CVC tests although the gap in performance between humans and the EIH system decreases as the highpass cutoff is increased from 200 to 2800 Hz.

The lower left panel compares human CV identification scores with the results from the three automated systems for CVC syllables subject to lowpass filtering. The machine identification scores are fairly constant at 90% as the cutoff frequency is decreased to 2000 Hz, falling to 80% for 1000 Hz lowpass filtering. The MFB and MFCC results largely overlap at cutoff frequencies above 2000 Hz, with the MFB results slightly superior to the MFCC results at lower cutoffs. The EIH front end consistently produces scores that are 5–10

Table 11
Human results for CVC syllables degraded by noise at $S/N = -7$ dB

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	31.5	19.2	24.6	3.4	0.5	0.5	15.3	0.5		2.5	1.5	0.5	203
t	0.5	89.0	4.7		1.0	0.5	2.1	0.5		1.0	0.5		191
k	14.2	26.5	35.8	1.2	1.2	2.5	14.2			2.5	1.9		162
b	1.9	3.7	3.7	22.2	10.5	5.6	8.0			38.3	4.9	1.2	162
d	1.7	3.5	2.3	6.9	32.4	8.7	8.1		0.6	23.1	12.1	0.6	173
g	3.5	2.5	3.0	8.5	8.5	28.6	5.5	0.5	1.0	25.6	12.6		199
θ	19.4	15.6	10.6	3.8	1.9	1.3	28.8	1.3	0.6	5.0	11.3	0.6	160
v					0.5		5.6	63.8	27.0	1.0		2.0	196
ð		0.6	1.3		0.6		7.5	8.2	80.5	1.3			159
s		2.0		9.8	7.2	6.5	5.9	1.3		51.0	11.8	4.6	153
ʃ	1.1	1.1		11.4	13.6	5.1	10.2			40.3	14.2	2.8	176
z					2.2	3.3	2.2	3.3	3.8	11.4	3.8	70.1	184
all	6.3	14.2	7.2	5.4	6.6	5.4	9.3	7.1	9.1	16.3	6.1	7.1	2118

See Table 3 for details.

Table 12
Machine results (MFB front end) on 1600 Hz highpass-filtered CVC syllables

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	70.7		14.3	7.5	2.0	4.8			0.7				147
t		95.2	2.1		2.1	0.7							146
k	4.1	4.7	82.4		0.7	7.4			0.7				148
b	38.5		0.7	44.8	6.3	6.3		2.8	0.7				143
d	1.4	20.9	0.7	1.4	68.9	5.4			1.4				148
g	4.9	7.0	25.9	6.3	4.2	51.0			0.7				143
θ	1.4						84.0	0.7	11.1	0.7		2.1	144
v	2.8			1.4	0.7	0.7	8.3	63.9	22.2				144
ð	2.8		0.7	2.1		0.7	7.6	8.3	77.9				145
s										91.1		8.9	146
ʃ											100.0		144
z										8.9		91.1	146
all	10.6	10.7	10.7	5.2	7.2	6.4	8.3	6.3	9.6	8.4	8.3	8.5	1744

See Table 3 for details.

percentage points below MFB and MFCC, except at the lowest cutoff frequency. At 500 Hz lowpass, EIH gave scores equivalent to MFCC, 10 percentage points below the MFB results.

The human CV results are generally inferior to the machine CVC results, but this may be reflect the difference between the relatively noiseless machine tests (+30 dB SNR) and the relatively noisy human tests (+12 dB SNR). For the highpass case, we have shown that the relatively noiseless human and machine CVC results are about 10–15 percentage points higher than the relatively noisy CV data. This is roughly the difference between the machine CVC results and the CV results for the lowpass case. However, the +12 dB additive flat-spectrum noise present in the CV tests should affect highpass filtering results more than lowpass filtering because the spectrum of speech generally falls with frequency above roughly 1000 Hz.

The lower right panel compares the results from the three automated systems with human identification scores for the additive noise conditions.

Table 13
Machine results (MFB front end) for CVC syllables degraded by additive speech-shaped noise at +10 dB SNR

	Recognized phone												N_i
	p	t	k	b	d	g	θ	v	ð	s	ʃ	z	
p	55.1	0.7	23.1	12.2	2.7	3.4	1.4	0.7	0.7				147
t	97.9	1.4		0.7									146
k	8.8	6.8	81.1			2.7	0.7						148
b	3.5	0.7		52.4	11.2	3.5	1.4	9.8	17.5				143
d	4.1	2.0	1.4	6.8	73.6	4.7		3.4	4.1				148
g	3.5	0.7	12.6	5.6	8.4	62.2		4.2	2.8				143
θ	4.2	0.7	5.6	1.4			82.6	2.1			0.7	2.8	144
v	3.5			10.4		0.7	9.0	48.6	27.1			0.7	144
ð	3.4		1.4	11.7	2.1	0.7	6.2	24.8	49.7				145
s		0.7								95.2		4.1	146
ʃ											100.0		144
z										6.2		93.8	146
all	15.4	1.1	10.6	8.4	8.3	6.4	8.4	7.7	8.4	8.5	8.3	8.5	1744

See Table 3 for details.

The MFB system is consistently better than the MFCC system except at 30 dB SNR, where the two points overlap. The EIH front end consistently yields the poorest identification scores. The performance difference between the EIH front end and the other front ends is not as large as for highpass-filtered speech, being more similar to the results for lowpass-filtered speech. The two sets of human recognition results largely overlap despite differences in the spectra of the noise: speech-shaped in the case of the CVC syllables and flat-spectrum in the case of the CV syllables and small differences in the number of items (12 CVCs and 16 CVs). Whereas for both of the filtering conditions the machine systems approached or exceeded human-level recognition rates, for speech degraded by additive noise machine recognition accuracy ranges from 15–30 percentage points lower than human accuracy.

4.3. Relative information transfer scores

Relative information transfer scores (Eq. (3)) were calculated for four divisions of the consonant set corresponding to the phonetic features of Voicing, Frication, Place, and Sibilance as shown in Table 1. As in the analysis of error rates, the three automated systems displayed similar patterns of Relative Information Transfer scores.

4.3.1. Highpass filtering results

Fig. 2 shows Relative Information Transfer scores for the highpass filtering condition. The upper left panel shows Relative Information Transfer scores for Sibilance, the upper right panel shows scores for Voicing, the lower left panel shows scores for Frication, and the lower right panel shows scores for the Place feature.

Human scores for the CVC and CV syllable sets are remarkably consistent, except for the case of Frication. This is not unexpected for this distinction which can be made on the basis of high frequency cues. The CV syllables were tested in the presence the +12 dB SNR flat-spectrum noise, which would be expected to resemble frication noise more the much weaker speech-shaped noise (at +35 dB SNR) used for the CVC tests.

We now compare Relative Information Transfer scores for humans on the CVC and CV databases with EIH, MFB, and MFCC results. For Sibilance, human performance is roughly constant at an identification score of 70% out to a cutoff frequency of 1.6 kHz. Machine performance is considerably higher than human, with scores for the MFB and MFCC systems being nearly perfect, and superior to that for the EIH system, which is nonetheless better than for humans. A similar pattern is seen in scores for Frication in the lower left panel, although for this feature machine

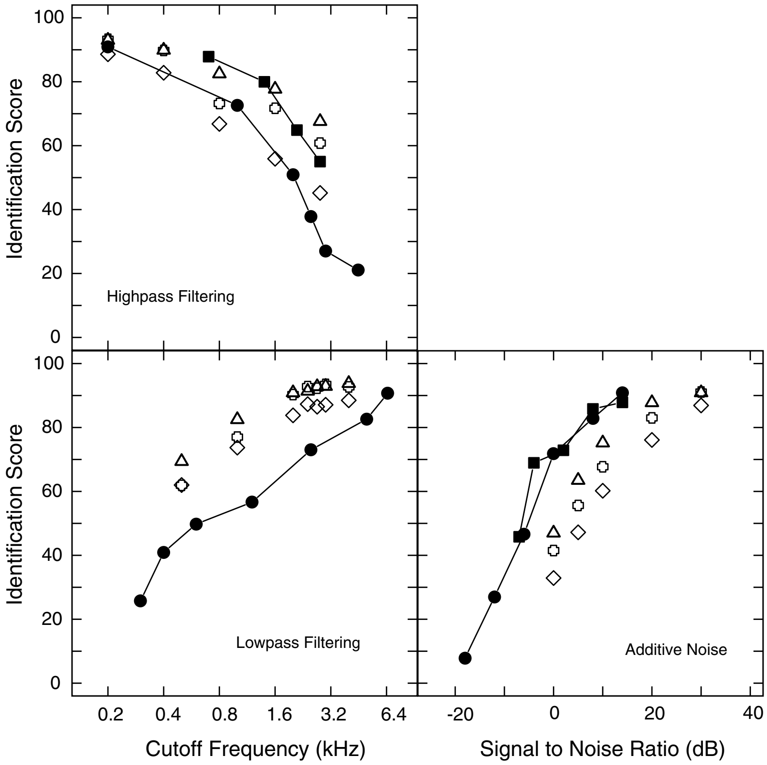

Fig. 1. Percent correct scores for humans and automatic systems on speech degraded by highpass filtering (upper left panel), lowpass filtering (lower left panel), and additive noise (lower right panel). The human results are marked by filled symbols, squares for our CVC results and circles for the Miller and Nicely (1955) CV results. The machine results on CVC utterances are marked by open symbols, diamonds for the EIH results, triangles for the MFB results, and crosses for the MFCC results. Note that the additive noise had a speech-shaped spectrum (at +35 dB SNR in the human filtering tests, at +30 dB SNR in the machine filtering tests, at varying SNR in the additive noise tests) in our study and a flat spectrum in the Miller and Nicely study (at +12 dB SNR in the filtering tests and at varying SNR in the additive noise tests).

performance is near perfect only for essentially unfiltered materials.

Scores for Voicing shown in the upper right panel, by contrast, are nearly the same for humans and the machine operating with the MFB and MFCC system, and the human superior to the EIH system for moderate degrees of filtering. Scores for Place (lower right panel) show a similar pattern of performance, with the MFB and MFCC scores higher by roughly 15 percentage points, and the EIH scores essentially the same as for humans.

Human scores are fairly consistent across the four features examined, reflecting the lack of pattern in their errors, which is seen in both the CVC and CV results. In contrast, machine systems show a greater ability to determine the Sibilance and Frication classes correctly compared to Place and Voicing classes, which introduces a pattern in the machine errors despite the similar number of phonetic confusions.

4.3.2. Lowpass filtering results

Fig. 3 shows Relative Information Transfer scores for lowpass-filtered speech. Scores for Sibilance, Frication, and Place are similar. The machine CVC scores are well above human CV scores, with the MFB scores generally above the MFCC scores, which are in turn above EIH scores. It should be noted, however, that the human CV tests were conducted in white noise at +12 dB SNR, while the machine CVC tests were conducted in speech-shaped noise at +30 dB

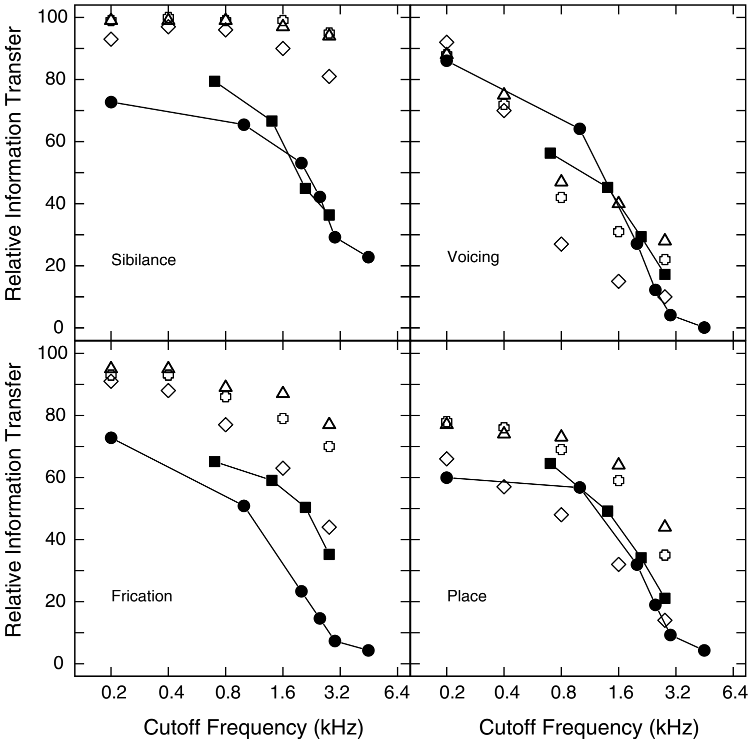

Fig. 2. Relative information transfer scores from the highpass filtering human results of Miller and Nicely (1955) on CV utterances, our human results on CVC utterances, and from our results on automated systems with CVC utterances. Symbols are the same as used in Fig. 1. Note that additive noise with a speech-shaped spectrum was used in our study (at +35 dB SNR in the human tests, at +30 dB in the machine tests) and with a flat spectrum was used in the Miller and Nicely study (at +12 dB SNR).

SNR. In spite of this difference, the human scores for Voicing were at roughly the level of the best machine systems. This is in contrast with the overall consonant recognition scores, seen earlier in the lower left panel of Fig. 1, which were superior for the machine systems. This shows for humans a greater robustness for recognizing Voicing over Frication and Sibilance. For both humans and machines, recognition of Place showed the lowest scores. In contrast to the highpass results, humans show varying abilities for recognizing each of the four features with lowpass filtering. Voicing is the most consistently recognized feature especially as the cutoff frequency is lowered.

4.3.3. Additive noise results

The final set of Relative Information Transfer score comparisons, seen in Fig. 4, is for speech degraded with additive noise. Scores for the Place and Frication features largely overlap for the two sets of human experiments. The differences between the two sets of human results are in the Voicing and Sibilance features. Tests using flat-spectrum noise (the CVs) resulted in higher Voicing RIT scores and lower Sibilance RIT scores when compared with tests using speech-shaped noise (the CVCs). This is partially due to the spectra of the respective noise maskers, with flat-spectrum noise having relatively more high-frequency energy than the speech-shaped noise.[5] Since the cues for Sibilance largely reside in the high-frequency portion of the speech spectrum, at a given SNR flat-spectrum noise is likely to be more

[5] Hant and Alwan (2003) have used a psychoacoustic-masking model to account for the difference in the effects of flat-spectrum and speech-shaped noise on the discrimination of plosive consonants.

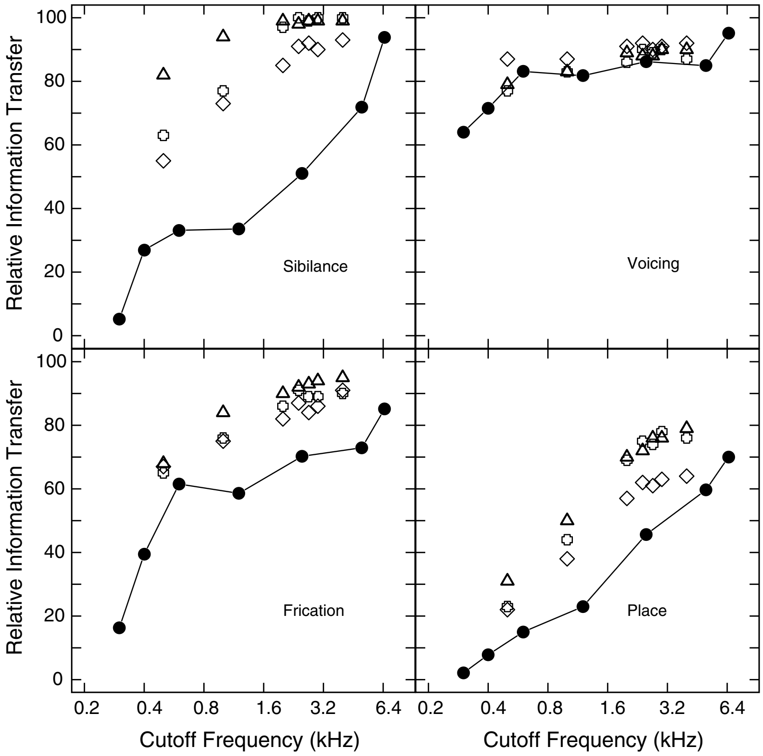

Fig. 3. Relative information transfer scores from the Miller and Nicely (1955) human CV data and from the automated systems on lowpass filtered CVC utterances. Symbols are the same as used in Fig. 1. Note that additive noise with a speech-shaped spectrum was used in our study (at +30 dB in the machine tests) and with a flat spectrum in the Miller and Nicely study (at +12 dB SNR).

deleterious than speech shaped noise. If Voicing is based on the low-frequency portion of the speech spectrum, the flat-spectrum noise would be less deleterious than speech-shaped noise at the same SNR.

Roughly similar patterns of machine performance are seen, with MFB scores higher than MFCC scores, which were in turn higher than EIH scores. Despite overall superiority of humans in noise, Sibilance RIT scores are comparable for human and machine systems using speech-shaped noise, and superior to those for flat-spectrum noise. Comparisons of the Frication and Place scores are consistent with the roughly 5–10 dB shift needed to match error rate scores of humans and the MFB system. Voicing scores, which are remarkably similar across the systems, suggest the need for a roughly 15–20 dB shift required to match MFB scores with humans. This reveals Voicing classification as a major weakness of the machine systems compared to humans. Voicing is the one phonetic feature we examined that has a temporal element to it, the intervocalic period, which distinguishes between its presence or absence. This temporal aspect may be a particular weakness of the HMM back end as opposed to any of the front end parameterizations we examined.

5. Summary and conclusions

The performance of our listeners on speech degraded by highpass filtering and additive noise is fairly consistent with that of Miller and Nicely (1955). This observation applies both to percentage correct scores and also to Relative Information Transfer scores. The differences in overall and RIT scores seem at least partially related to the fact that we performed our filtering tests in

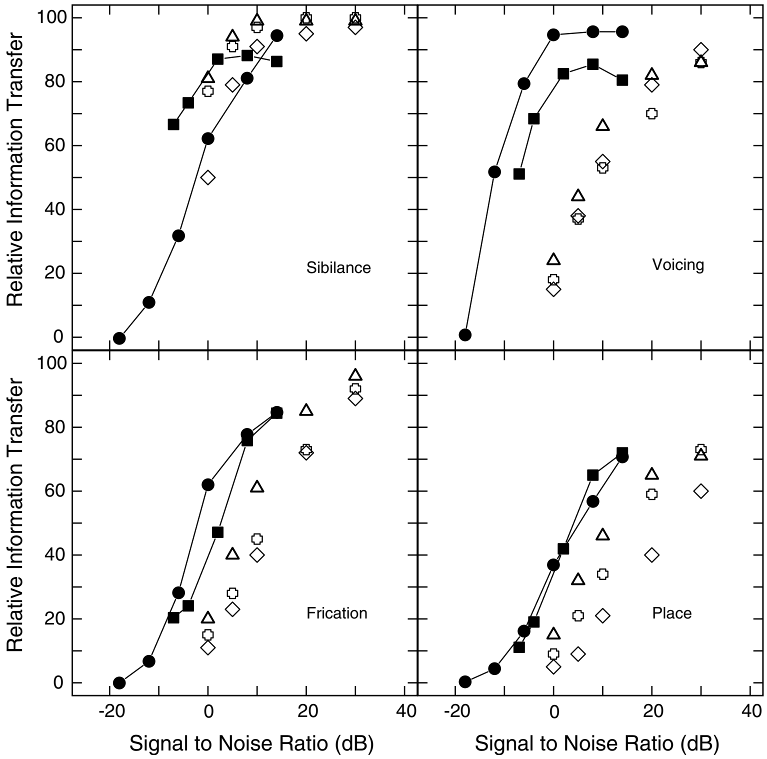

Fig. 4. Relative information transfer scores from the human CV results of Miller and Nicely (1955), from our human CVC results, and from our results on automated systems with CVC utterances, all degraded by additive noise. Symbols are the same as used in Fig. 1. Note that additive noise with a speech-shaped spectrum was used in our study and with a flat spectrum was used in the Miller and Nicely study. Relative information transfer scores from the highpass filtering human data of Miller and Nicely (1955) on CV utterances, our human data on CVC utterances, and from our results on automated systems with CVC utterances. Symbols are the same as used in Fig. 1. Note.

speech-shaped noise at +35 dB SNR and whereas Miller and Nicely tested in flat-spectrum noise at +12 dB SNR. Relative to flat-spectrum noise, speech-shaped noise has more power in the low frequencies and less power in the high frequencies.

For the machine systems tested, the highest levels of performance were exhibited by the MFB system, with scores for the MFCC system the same or very slightly inferior. Performance of the EIH system was inferior to that of the MFB and MFCC systems, both when evaluated in terms of overall identification scores and when evaluated in terms of the correct classification of individual features.

Machine scores reached or exceeded human levels for highpass filtered speech. For noisy speech, machine performance was significantly inferior to human performance, displaying roughly a 10 dB shift. For lowpass filtered speech, although this comparison is not on the same speech data, machine performance was comparable or superior to human performance. This difference was due in part to the difference in noise backgrounds used.

Machine response patterns were found to be similar for all three front ends tested. In particular, the use of Ghitza's Ensemble Interval Histogram (EIH) auditory-based front end did not improve ASR performance in degraded conditions. It also did not significantly alter the overall error patterns, which were more similar to the MFB and MFCC pattern than the human pattern.

On the CVC database, as the effectiveness of the speech cues is decreased through highpass filtering, the gap between human and machine abilities decreased or disappeared (top left of Fig. 1). This differentiates the effects of filtering from those of additive noise (bottom right of Fig. 1). In the case of additive noise, the gap in recognition accuracy between humans and machines increases as the speech is increasingly degraded.

The results reported here can be compared with those obtained by others. The Lippmann (1997) study cited earlier includes a comparison of human and machine recognition of spoken alphabet letters, with humans showing a 1% error rate and a neural network system showing a 5% error rate. Machine performance in this study seems roughly consistent with that result despite the vastly different approaches.

Jankowski's comparison of the EIH with a cepstral parameterization, like the Ghitza study, used systems trained on clean speech and tested in noise. On the isolated word recognition task, performance degraded mildly between +24 and +12 dB additive noise, rising from roughly 1% to 8% error rate, after which it degraded more quickly. Results not reported here (but see Sroka, 1998) showed an MFCC system trained on clean CVCs and tested across the range of noise degradations would require a roughly −20 dB shift in speech to noise ratio to match Jankowski's MFCC results. This disparity may arise in large part from their testing isolated word recognition on a 105-word database. Isolated words allow variability in many dimensions beyond that of the CVC nonsense syllables used here and could contain many cues that would not have been available to the systems being tested on consonant recognition.

Analysis of confusion matrices shows differences in the error patterns between humans and machines. In highpass conditions, machines make relatively fewer Frication and Sibilance classification errors while humans make relatively fewer Voicing classification errors. Place classification scores are comparable. The lack of pattern in human errors across these features for highpass filtered conditions was not seen in tests of ASR systems for the same conditions. The differences in response patterns suggest that the human subjects and the ASR systems were relying on different properties of the acoustic speech waveform for recognition despite the overall performance level similarities. The high information transfer scores for various features indicates that even in these filtered conditions cues persist, of which humans are not taking advantage, that allow feature determination at levels above those displayed by humans, and that machines can take advantage of those cues if trained in that condition.

Our findings are consistent with Ghitza's report of large differences between the results obtained from humans and machines in noise. Ghitza's study, like Jankowski's, used systems trained on clean speech and tested in noise. Unfortunately Ghitza does not report percent correct scores for identification of consonants. however, all of the features he examined at +10 dB SNR showed greater than 10% error rates for the EIH except one, while human error rate, if averaged, would be roughly 6%. The one case where the EIH error rate was less than 10% was for correctly classifying a Non-Sibilant sound, which showed minimal errors. Since Ghitza reports separate scores for classification of Sibilant sounds and Non-Sibilant sounds (Sibilance classification of a Sibilant compared to Sibilance classification of a Non-Sibilant), and since the correct classification of Sibilant sounds showed roughly a 40% error rate, it seems that in his study recognition was biased toward classifying sounds as Non-Sibilants. Averaging the error rates for Sibilants and Non-Sibilants leads to roughly a 20% error rate. Comparison with our implementation of the EIH is made difficult because Ghitza used a Two-Alternative Forced Choice approach (taking a speech sample and only comparing two HMMs at a time, one HMM for the correct phone sequence and one for a phone sequence that differed by only one consonant feature), because the testing in degraded conditions was not matched with training in those conditions, and because Ghitza used speech that was lowpass filtered to 3600 Hz. Still, at +30 dB SNR Ghitza showed an average feature error rate of roughly 10%, which is consistent with the EIH results reported here for 4 kHz lowpass filtered speech.

It is possible that cues being used by the machine systems are only valid for the set of consonants tested in the study, and that humans ignore them because they become less useful when the complete phone set of English is considered.

Furthermore, the limited number of speakers (2) and the Degradation-Specific Training could have allowed the machine systems to learn cues in particular degraded conditions that are not robust to additional speakers or across the degraded conditions. This could explain the machine results in highpass filtered conditions which actually outperformed human listeners. The training approach we describe is of limited value commercially, because it requires detailed knowledge of the operational environment during training and assumes consistent signal degradation. However, the fact that a machine system is able to recognize at these levels shows that there are cues present that allow recognition at levels above those displayed by humans, and that the HMM training was able to learn those cues in these favorable training and testing conditions. This was not seen for the additive noise tests.

It is also possible that the cues being used by the machine systems are used by some listeners but not those tested. This is consistent with the double-weak theory of speech perception (Nearey, 1997), which proposes articulatory targets are developed within a language in order to produce a reliable (but possibly variable) set of cues that can include redundancies. This would allow recognition to use varying cues depending on the speech, the listener, the talker, and the environment in which the speech was occuring.

In additive noise conditions, humans outperformed the ASR systems. Human and machine error patterns were more similar than in filtering conditions, especially when overall performance levels were matched, though not as similar as the response patterns between the different ASR systems. As in the filtering conditions, machines displayed relatively good performance for the Sibilance classification and relatively poor performance for the Voicing classification (equivalent to a difference of 15–20 dB in speech to noise ratio).

Overall, these results motivate continued work at the phone recognition level for improving speech recognition and understanding systems. Work in recent years has shown that the introduction of lexical, syntactic, and other higher-level speech analysis can improve automated speech recognition. However, until classification of the basic speech sounds can reach human levels, we cannot know that continued improvements at introducing these higher-level analyses will enable human-level recognition.

Our results may motivate directed attempts to augment traditional parameters in ways that will improve their ability to make the Voicing distinction in order to produce a more noise-robust system. Attempts to do frame-based classification of phonological features using neural networks have shown promising results (King and Taylor, 2000). Using HMMs to do word recognition based on the outputs of neural networks performing frame-based phonological feature classification has shown performance levels comparable to traditional approaches (Kirchhoff and Bilmes, 1999). Further, augmenting a traditional frame-based front end parameterization with frame-based phonological feature values based on analysis of the speech signal was found to improve performance (Kirchoff, 1998) over the traditional front end alone. A more directed application of phonological classification based on identified phonological feature weaknesses may yield finer resolution into potential improvement by indicating a subset of phonological features that are consistently misrecognized by machine systems. Phonological feature analysis could then be directed to just those features that are typically misrecognized rather than attempting to recognize a complete feature set, minimizing the number of parameters (and therefore the computational cost) of an improved recognizer. In this view, degrading speech through filtering or additive noise would affect the various cues differently, possibly creating patterns of errors as performance degraded, and would affect human listeners differently depending on the extent to which the particular cues they relied on were robust to the signal degradations.

The differences in machine and human error patterns and robustness to manipulations of the speech signal supports the interpretation that they are using different acoustic cues in the recognition process. This motivates additional work to determine the specific acoustic cues that humans use to recognize basic speech sounds more robustly than machines. Knowledge of the cues used by humans could be used to direct future research by identifying the cues that human test results show are present and robustly detectable.

As an example, the disparity in Voicing classification scores may arise from the temporal nature of one of the key perceptual cues used by humans, the intervocalic period (Klatt, 1975; Stevens, 1980, 1992). The Hidden Markov Models as used in this research do not provide a means for modeling a Gaussian distribution of a durational cue like intervocalic period to make it useful for classification. Thus, the HMM pattern recognition approach is unable to leverage one of the cues known to be relied upon by humans.

To address this limitation of HMMs, a subphonetic feature detection stage (as opposed to the phonetic feature classification approaches described above) could be designed specifically to recognize a cue like intervocalic period prior to the HMM recognition stage. The duration of the intervocalic period could then be used to augment the traditional parameters for the HMM. This approach to improving ASR attempts to decrease the performance difference between humans and machines by making available an additional piece of information (the intervocalic period that HMMs are otherwise unable to model), selected for its discriminative ability for Voicing in Plosives, an identified weakness of the automated systems.

References

ANSI (1997). Methods for the Calculation of the Speech Intelligibility Index, ANSI S3.5-1997. American National Standards Institute, New York.

Bustamante, D.K., Braida, L.D., 1986. Multiband compression limiting for hearing-impaired listeners. J. Rehab. Res. Dev. 24 (4), 149–160.

Bratakos, M.S., Reed, C.M., Delhorne, L.A., 2001. A single-band envelope cue as a supplement to speechreading of segmentals: a comparison of auditory versus tactual presentation. Ear Hearing 22 (3), 225–235.

Bustamante, D.K., Braida, L.D., 1987. Principal-component amplitude compression for the hearing impaired. J. Acoust. Soc. Am. 82 (4), 1227–1242.

DeGennaro, S., Braida, L.D., Durlach, N.I., 1986. Multichannel Syllabic Compression for Severely Impaired Listeners. J. Rehab. Res. Dev. 23, 17–24.

Dix, A.K., Braida, L.D., 2002. Effect of noise on the recognition of CVC consonants. Unpublished manuscript.

Ebel, W.b., Picone, J., 1995. Human speech recognition performance on the 1994 CSR spoke 10 Corpus. In: Proc. of the Spoken Language Systems Technology Workshop, pp. 53–59.

Ghitza, O., 1993. Adequacy of auditory models to predict human internal representation of speech sounds. J. Acoust. Soc. Am. 93 (4), 2160–2171.

Ghitza, O., 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans. Speech Audio Process. 2 (1), 115–132.

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

Hidden Markov Model Toolkit, CUED-HTK, 2004. University of Cambridge, UK (http://htk.eng.cam.ac.uk/).

Jankowski Jr., C.R., Vo, H.-D.H., Lippmann, R.P., 1995. A comparison of signal-processing front ends for automatic word recognition. IEEE Trans. Speech Audio Process. 3 (4), 286–293.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Computer Speech Language 14, 333–353.

Kirchoff, K., 1998. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In: Proc. 5th Intl. Conference of Spoken Language Processing, Australian National University, Sydney, Aus., Decemeber.

Kirchhoff, K., Bilmes, J.A., 1999. Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values. In: Proc. 1999 IEEE Intl. Conference on Acoustics, Speech, and Signal Processing, (IEEE Cat. No. 99CH36258).

Klatt, D.H., 1975. Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. J. Speech Hearing Res. 18, 686–706.

Lippmann, R.P., 1997. Speech Recognition by Machines and Humans. Speech Commun. 22 (1), 1–15.

Lippmann, R.P., Braida, L.D., Durlach, N.I., 1981. A study of multiband amplitude compression and linear amplification for persons with sensorineural hearing loss. J. Acoust. Soc. Am. 69 (2), 524–534.

Mathworks. Inc. MATLAB Versions 4.8–5.2. The MathWorks, Inc., Natick MA, copyright 1984–2004.

Miller, G.A., Nicely, P.E., 1955. An Analysis of Perceptual Confusions Among Some English Consonants. J. Acoust. Soc. Am. 27 (2), 338–352.

Nearey, T.M., 1997. Speech perception as pattern recognition. J. Acoust. Soc. Am. 101 (6), 3241–3254.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.

Ronan, D.E., Dix, A.K., Shah, P., Braida, L.D., 2004. Integration of Acoustic Cues for Consonant Identification across Frequency Bands. J. Acoust. Soc. Am. 116 (3), 1749–1762.

Slaney, M., 1998. Auditory Toolbox, Version 2. Technical Report #1998-10, Interval Research Corporation. http://rvl4.ecn.purdue.edu/malcolm/interval/1998-010/.

Sroka, J., 1998. Consonant recognition by humans and machines. Ph.D. Thesis, Massachusetts Institute of Technology, Division of Health Sciences and Technology, Cambridge, MA, September.

Stevens, K.N., 1980. Acoustic Correlates of Some Phonetic Categories. J. Acoust. Soc. Am. 68 (3), 836–842.

Stevens, K.N., Blumstein, S.E., Glicksman, L., Burton, M., Kurowski, K., 1992. Acoustic and Perceptual Characteristics of Voicing in Fricatives and Fricative Clusters. J. Acoust. Soc. Am. 91 (5), 2979–3000.

Winer, B.J., 1971. Statistical Principles in Experimental Design. McGraw-Hill, New York.