2pPPb28. Assessing the contribution of spectral cues to recognition of frequency-lowered consonants

Kelly Fitz*, Christophe Micheyl, Susie Valentine and Tao Zhang

*Corresponding author's address: Signal Processing Research Department, Starkey Hearing Technologies, 6600 Washington Ave. S., Eden Prairie, MN 55344, kelly_fitz@starkey.com

Commercially available strategies for restoring audibility of critical high frequency cues to patients with severe high frequency hearing loss translate information from high-frequency regions with unaidable hearing to lower-frequency regions with aidable hearing. Methods for synthesizing lowered spectral features, rather than generating them from the signal, have been proposed, though no commercially available hearing aid uses such a method. We assessed consonant discrimination under three configurations of a spectral feature synthesis method intended for use in frequency lowering. Lowered consonants were rendered using one or two narrowband noise components presented in a low frequency region with aidable hearing. Different configurations conveyed different spectral cues intended to distinguish among lowered consonants. In a short pilot study, preliminary analysis found no significant difference in consonant matching accuracy between the different configurations, suggesting that listeners were not making use of additional spectral cues when they were available. However, there were some observable trends in the data, as well as open questions about the possible impact of training and acclimatization on listener performance. We will present the findings of a more extensive study to determine whether listeners with training can make use of enhanced spectral cues to distinguish among frequency-lowered consonants.

Published by the Acoustical Society of America through the American Institute of Physics
OBJECTIVES

High frequency sounds are critical to speech intelligibility, with a substantial portion of audible speech cues occurring at frequencies higher than 3 kHz. Restoration of audibility for these high-frequency speech cues is often constrained by the power available in the hearing aid, by the amount of gain that can be applied without introducing feedback, and by the extent of the cochlear damage.

Commercially available strategies for restoring audibility of critical high frequency cues to patients with severe high frequency hearing loss translate information from high-frequency regions with unaidable hearing to lower-frequency regions with aidable hearing. Methods for synthesizing lowered spectral features (rather than generating such features from the speech signal itself) have been proposed. However, no commercially available hearing aid uses such a method. In one recent example of such a method, Kong and Mullangi (2012a) describe a technique that synthesizes spectral differences in frequency-lowered fricative consonants according to a phoneme-classification system, with the goal of reducing place-of-articulation errors, which represent the most common form of consonant-recognition error in hearing-impaired listeners (Dubno et al., 1982).

It has been suggested that listeners with hearing loss may need time or training to learn to use the remapped spectral cues provided by frequency-lowering technologies (Robinson et al., 2009; Simpson, 2009). Moreover, even though intelligibility of fricative consonants may improve with frequency lowering, place-of-articulation errors remain common among patients who use these technologies in hearing aids (Kong and Mullangi, 2012b). It is therefore of interest to identify efficacious frequency-lowering strategies that facilitate the learning of the lowered spectral cues and reduce the incidence of place-of-articulation confusions.

This manuscript describes a study in progress that is intended to determine whether, with training, listeners can make use of enhanced spectral cues to discriminate among frequency-lowered consonants. The outcome of this study will provide and indication of the importance of preserving and rendering spectral differences with frequency lowering, and of the ability of listeners with hearing loss to learn to use spectral cues in frequency-lowered consonants to discriminate among consonants differing only in place-of-articulation.

BACKGROUND

In a pilot study, we assessed consonant discrimination under three different frequency-lowering treatments employing synthesized features, the treatments differing in their preservation of spectral contrasts among the synthesized cues. Subjects were presented vowel-consonant (VC) nonsense tokens and asked to match consonants across different vowel contexts. In a three-interval, two-alternative forced-choice (3I-2AFC) task, two test intervals (target and non-target, in random order) contained VCs differing only in consonant (e.g. /is/ and /if/), and a reference interval contained a VC with the same consonant as the target interval, but differing in vowel (e.g. /as/). Correct response required that the subject match consonants across vowel contexts. Success in this task requires that lowered features for a particular consonant sound similar enough to match across utterances and vowel contexts, and distinct from features for other consonants.

We employed a consonant-matching task, rather than a consonant-identification task, because we assumed that listeners would need a considerable amount of exposure to successfully learn to associate the new cues with specific consonants when these were presented in the absence of (linguistic) contextual cues. Even cues that are very distinct and easily learned may not immediately be identified with the consonant that they are intended to represent. For example, it is common to misidentify a frequency-lowered /s/ as /sh/ (a confusion of
place-of-articulation) (Robinson et al., 2009). The matching task was designed to explicitly exercise the ability to form classes of lowered consonant sounds, and to recognize the cues for one class as sounding more similar to each other than to members of another class, without relying on the listener’s ability to immediately form associations with particular consonants. We tested consonants that shared common manner and voicing, but differed in place-of-articulation, with the expectation that place classes would be more perceptually distinct under treatments that provide more spectral information in the lowered cues than under treatments that provide less spectral information in the lowered cues.

If listeners make use of spectral cues to discriminate among the different frequency-lowered consonants, then we expect that performance on the matching task would be better under the treatments that present greater spectral contrasts among the lowered consonants. A treatment employing a phoneme classifier to render spectral contrasts that are consistent across utterances, as recommended by Kong and Mullangi (2012a), might further improve performance. Alternatively, it may be that the primary benefit afforded by frequency lowering is due to the audible temporal features of the processed speech, and does not depend on preservation or rendering of spectral contrasts.

Preliminary analysis found no significant difference in consonant-matching accuracy between treatments differing in their preservation of spectral contrasts, suggesting that listeners were not making use of additional spectral cues when they were available. However, there were some observable trends in the data, as well as open questions about the possible impact of training on listener performance. Previous research suggests that listener’s ability to use frequency-lowered speech cues improves with training (Füllgrabe et al., 2010; Robinson et al., 2009; Simpson, 2009), and it is possible that listeners might learn, with training, to take advantage of the enhanced spectral cues provided by some of the treatments in our study. Therefore, one objective of the current study is to determine whether more exposure to the frequency-lowered speech, through training, will enable listeners with hearing loss to use spectral information to better differentiate fricative consonants.

METHODS AND MATERIALS

Participants

Participants in this study will have steeply sloping sensorineural hearing loss, and audiometric thresholds that satisfy the following criteria:

1. Hearing loss must be worse than 65 dB HL at one or more frequencies below 4 kHz, and at all frequencies above 4 kHz.
2. Hearing loss must be 50 dB HL or better at all frequencies below 1.5 kHz.
3. For at least one octave, the slope of the audiogram must equal or exceed 25 dB HL per octave.

Both ears must satisfy all three inclusion criteria. If asymmetry exists between ears that both meet the inclusion criteria, the better ear (higher average of pure tone thresholds at 500, 1000, and 2000 Hz) will represent the test ear and the worse ear the non-test ear. Using the better ear as the test ear will increase the likelihood that masking the non-test ear will prevent it from responding to test stimuli.

All participants will complete a qualification session to establish the suitability of the frequency-lowering parameter configuration. Participants for whom an efficacious configuration cannot be found will not be tested further.
Participants will complete a pre-test, a training phase, consisting of four graduated training sessions, and a post-test. The entire sequence will be conducted over four sessions on separate days within a single week. Participants will complete three sequences, one for each treatment, in separate weeks with at least two weeks separating sequences. Participants in a separate control group will complete only the pre- and post-tests, separated by three or four days, with no training in between. Data from this control group will allow us to confirm the impact of training independently of the impact of repeated exposure to the task.

Stimulus Presentation and Compensation for Hearing Loss

Test stimuli will be presented monaurally under Sennheiser HD-600 headphones, which are open-air diffuse-field-compensated circumaural headphones. Nominal speech presentation levels will be between 55 and 65 dB SPL, before amplification, which will produce levels near 90 dB SPL (i.e., approximately 30 dB gain). To minimize cross-hearing in the non-test ear, a broad-band speech masker will be applied to the non-test ear at 50 dB SPL based on interaural attenuation values near 40 dB for circumaural headphones (Brännström and Lantz, 2010).

To compensate for hearing loss, linear gain shaping will be applied to the stimuli using NAL-NL2 (Keidser et al., 2011) target insertion gains for speech at the targeted presentation level of 60 dB. Linear gain is chosen to avoid the potential confound of interaction of the frequency-lowering treatments with dynamic behavior of a multichannel compressor.

Treatments

Test stimuli will be processed offline according to each of three frequency-lowering algorithms. All three methods synthesize lowered spectral features, rather than transposing energy from an input signal. To isolate the effects of spectral-feature synthesis from the effects of the spectral feature identification (or analysis), all three treatments employ a single, “oracle” analysis to identify high-frequency consonants for all three treatments. Voiceless fricative consonants in each speech stimulus are identified, by inspection and audition, and labeled according to temporal extent (beginning and end time) and phoneme (/f/, /θ/, /s/, or /ʃ/).

Spectral analysis is performed only in the labeled consonant regions. Within these regions, the signal is highpass filtered at 2 kHz, and power-spectrum analysis is performed on overlapping 125 ms frames. Within each frame, the power is estimated in the neighborhood of each spectral peak, and the two peaks having the highest power are selected to represent the frame. Peaks are connected from frame to frame to form “track”, each track representing the time-frequency evolution of a ridge in the short-time power spectral surface. The parameters (frequency and amplitude, or power) of these tracks are used to guide the synthesis of frequency-lowered features for the voiceless fricative consonants.

Adapting a technique from musical sound synthesis, spectral features are rendered by additive synthesis of tracked spectral peaks, where the additive components are modulated copies of a single prototype narrowband noise (Fitz et al., 2003). Lowered consonants will be rendered using one or two narrowband noise components presented in a low-frequency region with aidable hearing. Different configurations convey different spectral cues intended to distinguish among lowered consonants. This synthetic frequency-lowering method gives a high degree of control and predictability of the translated feature spectra, and is practical for implementation in a digital hearing aid.

Treatment 1 - One Component: The first treatment combines the power from the (at most) two tracked peaks in the consonant regions into a single lowered feature (narrowband-noise components), at a fixed frequency.

Treatment 2 - Two Components: The second treatment preserves the power and frequency
relationships of the (at most) two tracked peaks in the consonant regions, and renders two noise components having power equal to that of the tracked peaks, and center frequencies lowered by a constant factor. This method produces (at most) two components having center frequencies that vary with the frequencies of spectral peaks in the analyzed speech.

**Treatment 3 - Two Components With Classification:** The third treatment uses an algorithm derived from an algorithm proposed by (Kong and Mullangi, 2012a) to synthesize different spectra for different classes (defined by place-of-articulation) of non-sonorant consonants. In this treatment, each class of consonant is rendered according to a different “recipe”, and only the power in the tracked peaks (not the frequency) is considered in the synthesis. Consonants in the dental/labio-dental group (/f/, /v/, /th/, /dh/) are synthesized by two components at fixed frequencies, having power equal to the two tracked peaks. Consonants in the alveo-palatal group (/sh/, /zh/) are synthesized using only the lower-frequency component, combining the power from the two tracked peaks. Consonants in the alveolar group (/s/, /z/) are synthesized using only the higher-frequency component, combining the power from the two tracked peaks.

The time-varying frequency and amplitude parameters of the synthesized components for the word “forks” are plotted in Figure 1. Note that, under treatments 1 and 3, alveo-palatal and alveolar consonants are rendered using a single fixed-frequency noise component, so depending on the choice of component frequencies in those two treatments, consonants from one group or the other may be synthesized identically under these two treatments. Since the /s/ sound generally produces the highest-frequency spectral peak, and with the goal of configuring the processing to produce the highest-frequency-lowered features near the upper end of the aidable frequency range, we will configure the single component in treatment 1 to have center frequency equal to that of the higher frequency component in treatment 3, so that alveolar consonants (the key consonants in the S-test) will produce the same lowered cues under these two treatments.

Treatment 2 is always distinct, because the center frequencies of the synthesized components vary according to the spectrum of the analyzed speech. However, we will attempt to map the range of likely peak frequencies onto a range similar to the one spanned by the components in treatment 2.

We employ fixed-frequency components at 500 and 1250 Hz, in treatments 1 and 3, and in treatment 2, scaled peak frequencies by 1/5, thereby mapping peak frequencies between 2500 and 7000 Hz onto a range spanning 500 to 1400 Hz. For subject who perform poorly in the qualification phase (S-Test), we apply an alternate parameter configuration having fixed frequency components at 600 and 1600 Hz, in treatments 1 and 3, and in treatment 2, scaled peak frequencies by 1/4, thereby mapping peak frequencies between 2400 and 7000 Hz onto a range spanning 600 to 1750 Hz.

**FIGURE 1:** Plots of time-varying frequency and amplitude parameters applied by each of the three treatments for the word “forks”.
Speech Intelligibility Measures: S-Test

The S-test (Robinson et al., 2007) will serve as a baseline measure of benefit accrued by the frequency-lowering treatments. The S-test evaluates important high-frequency word-final cues that provide important linguistic information for spoken English. The stimulus set consists of four utterances of each of 24 word pairs spoken by a female British English speaker, with the pairs consisting of words differing only in the presence or absence of word-final /s/ or /z/ (e.g. “book” - “books”). In each trial, the subject hears a single word, and indicates whether they heard the singular or plural, using a graphical computer user interface developed for this purpose.

The 96 word pairs are divided into two test lists, each consisting of two instances of each of the 48 singular and plural words. All three treatments are applied to each of the two lists, making six treated word lists, plus two untreated lists that are used to establish baseline performance, against which the benefit due to treatment is measured. Each subject will be tested on each of the eight word lists in a single session. Additionally, one short (24 word) practice list consisting of singular and plural unprocessed stimuli drawn from the entire 192 word corpus will be completed first, to familiarize the subject with the task. Test order will be randomized to control for ordering effects, but controlled to prevent the same treatment being presented in consecutive tests. Experience suggests that each 96-word list can be completed in approximately ten minutes.

We apply a low level (20 dB SNR) speech-shaped masking noise to conceal voice-offset cues that might affect performance and interfere with our ability to measure benefit accrued due to the frequency-lowering treatment. We analyze S-test performance using the sensitivity index, $d'$, which accounts for both hit rate and false alarm rate, and provides a metric on the difference in response to the “signal-present” and “signal-absent” stimuli, where the “signal” is the word-final /s/ or /z/. A $d'$ score of 0 indicates chance performance (the subject is guessing), and a score of 1.0, as a rule of thumb, is taken to indicate that the subject is just able to detect the signal (corresponding to about 75% correct responses on a 2AFC task), and their performance is unlikely to have occurred by chance. A score of 3.0 is generally considered good performance, corresponding to about 98% correct responses on a 2AFC task.

Presentation level during the test session will be determined by evaluating the performance on a single unprocessed practice word list, consisting of one instance of each of the 48 singular and plural words, drawn from the same stimulus pool as the test lists, presented at a nominal level of 60 dB SPL. If performance is too low ($d'$ lower than 0.6), the nominal presentation level for the tests lists will be raised to 65 dB SPL. If performance is too high ($d'$ higher than 1.1), the nominal presentation level for the tests lists will be lowered to 55 dB SPL.

Compared to the consonant matching task that is the focus of this study, the S-test is a simpler task that requires only that the frequency-lowering treatment be audible to participants. If the processing is not audible, then subjects will clearly not be able to make use of the spectral cues. Therefore, we can use the S-test results to confirm that our frequency-lowering algorithms are appropriately configured for the subject.

Speech Intelligibility Measures: Consonant matching

Task

In a three-interval two alternative force choice (3I-2AFC) task, participants will hear a sequence of three nonsense syllables. The first, reference interval, containing the target consonant, is followed by two test intervals, only one of which contains the target consonant. The reference stimulus will differ from the test stimuli by the vowel context, and the two test stimuli will differ only by consonant (same talker, position, and vowel). All three stimuli will be
processed using the same frequency-lowering strategy. The participant will be asked to identify
the test interval that contained the same consonant as the reference interval using a graphical
computer interface. For example, in a trial consisting of a reference stimulus /is/ and test
stimuli /as/ and /af/, the correct response is /as/.

Testing and training will tasks will be identical, only the stimuli will differ.

Stimuli

The stimuli will be drawn from the Non-sense Syllable Test (NST). The consonant tokens
will include the four unvoiced fricative consonants differing in place-of-articulation, /f/ (labio-dental), /th/ (dental), /s/ (alveolar), and /sh/ (alveo-palatal), and three different vowel
contexts, /a/, /i/, and /u/. Consonant-vowel (CV) and vowel-consonant (VC) utterances from
two talkers (male and female) will be used. Forty-eight (48) stimuli provide one instance of each
consonant, vowel, position, talker combination. Nominal speech presentation level will be the
same performance-adjusted presentation level used in the S-Test (see above). Stimuli will be
presented in quiet (no background noise in the test ear), and, as described in Section 3.4, a
broad-band speech masker will be applied to the non-test ear at 50 dB SPL.

Each trial in the testing phase presents a reference interval pairing one of the four
consonants (e.g. /s/) with one of three vowels (e.g. /as/, /is/, /us/). The reference is paired
with two different target intervals (in separate trials) differing only in vowel (e.g. /as/-/is/ and
/as/-/us/). Each reference-target pair is matched with three different nontarget intervals (in
separate trials) differing from the target only in consonant (e.g. /as/-/is/-/if/ and
/as/-/is/-/ish/ and /as/-/is/-/ith/). Thus, there are three times two times three equals 18
different reference-target-nontarget trios per consonant. Each trio exists in two talker (male
and female) times two placement (CV and VC) forms, so the total number of unique stimulus
combinations is four consonants times 18 trios times four talker-placement settings equals 288
combinations. From these combinations, a test set of 144 trials will be randomly selected,
balanced to include one male and one female presentation, and one CV and one VC form for
each reference-target-nontarget trio. The order of presentation of the target and non-target
intervals will be randomized on each trial, but only one order will be presented in a run.

Test sessions will be comprised of two runs of the selected 144 trials in random order, with a
short rest between runs. Each subject will complete the test once at the beginning of the week
(pretest) and once at the end of the week, after training (post test). The same 144 trials will be
presented in a different random order in both pre- and post-training tests. Subjects in the
control group will not receive training, but will complete the pre- and post-training tests in
sessions separated by three or four days (like the trained group).

In the initial pilot study, each of the trios was presented by two talkers, but using only VC
placement forms (because previous studies such as Dubno et al. (1982) have found this form to
be the more challenging for listeners with hearing loss). This resulted in a total of 144 trials,
presented in a single run, which was typically completed in approximately 30 minutes.

Graduated training regimen

Training runs will consist of the same consonant matching task, but the stimulus sets and
the composition of the stimulus trios presented in each trial will be manipulated to make the
task easier in the earlier sessions, and gradually build in complexity until the final training run
uses the same trial composition as the test runs. There will be four training sessions, completed
on different days. The post-training test will be administered following the final training
session.
DATA ANALYSIS

In considering the effect of treatment, we are primarily concerned with two questions:

1. Does the addition of spectral cues give listeners with training a significant advantage in distinguishing and classifying frequency-lowered fricatives (treatments 2 and 3, versus treatment 1)?

2. Does a method of adding consistent spectral cues based on (oracle) phoneme recognition/classification give listeners with training a significant advantage in classifying lowered fricatives relative to a method based solely on spectral analysis of the stimulus (treatment 3 versus treatment 2)?

Data were analyzed using a combination of frequentist and Bayesian statistical methods. Preliminary analysis of variance with repeated factors of session, consonant, and treatment revealed no significant effect of treatment on consonant matching accuracy, and a significant interaction of treatment with consonant pair only for the /f/-/sh/ pair, and only when a subgroup of two low-performing subjects (identified by Kruskall-Wallis ANOVA) were removed from the analysis.

Subsequently, the data of the first experiment were analyzed using a Bayesian model. The data were in the form of correct-response counts. The simplest, and most commonly assumed, statistical model for such count data is the binomial distribution. This distribution is entirely characterized by two parameters, which correspond to the number of trials and to the probability of “success” (i.e., of a correct response). Accordingly, the number of correct responses (out of 12) for a given subject (j), treatment (k), consonant pair (c), and consonant-pair order (r), \( n(j,k,c,r) \), was modeled as a binomial random variable with a correct-response probability, \( p(j,k,c,r) \), and parameter \( n = 12 \) (the number of trials).

Following another common approach in the statistics literature, the probit (i.e., inverse cumulative standard normal distribution function) of the probability, \( p(j,k,c,r) \), was modeled as a linear combination of effects. Specifically, we considered a model with two main effects related to the subject and the consonant pair (taking into account the order), respectively, and an interaction term, which depended on the combination of treatment and consonant pair. Formally, we set

\[
\logit[p(j,k,c,r)] = \mu + ke(k,c,r) + ce(c,r) + je(j),
\]

where \( \mu \) is a constant term, \( ke(k,c,r) \) is a a treatment effect, \( ce(c,r) \) is a consonant-pair effect, and \( je(j) \) is a subject-specific effect.

Consistent with the Bayesian framework, all variables in the model were assigned prior distributions. The variables, \( ke, ce, \) and \( je \) were assigned vaguely informative independent \( \mathcal{N}(0,1) \) priors, and \( \mu \) was assigned a \( \mathcal{N}(0,10) \) prior. The constant term, main effects, and interactions were inferred using Markov-chain Monte-Carlo methods. Inferences were in the form of posterior distributions, from which estimates of central tendency (mean, median, mode) and “credible regions” (the Bayesian equivalent of confidence intervals) could be computed. Although 95% credible regions for differences between treatments taken pairwise (e.g., treatment 2 minus treatment 1) always contained zero, for the consonant pair, /th/-/s/ the 50% credible region for the difference between treatment 3 and treatment 1. This preliminary outcome suggested that, for this consonant pair, a beneficial effect of treatment 3 over treatment 1 might be observed with more subjects, and prompted the follow-up experiment described above.

The data of the present experiment will be analyzed using an extended version of the Bayesian model described above, modified to accommodate the additional “test session” factor.
(pre- versus post-training). The latter will allow us to examine the effect of training on consonant matching performance, to answer the questions:

1. Did listeners with training perform differently at the end of the week (on the post-test) from the beginning of the week (on the pretest)?

2. Did listeners with training perform differently (or did their performance evolve differently) from listeners in the control group that received no training?

If the training is effective, we expect to see that post-test scores are higher than pretest scores among the trained listeners, and that the difference in scores is greater than any difference shown by listeners in the control group.

Finally, each of the test data sets (pre- and post-training) will be presented twice (two repetitions), to allow us to assess repeatability of the subjects' responses, by examining the strength of response correlation within stimulus trios between repetitions.

Data collection for the present experiment is in progress, with results to be reported at the meeting.

REFERENCES


