Measuring time-frequency importance functions of speech with bubble noise[a]

Michael I. Mandel[b]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

Sarah E. Yoho and Eric W. Healy
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

(Received 2 March 2016; revised 1 September 2016; accepted 12 September 2016; published online 13 October 2016)

Listeners can reliably perceive speech in noisy conditions, but it is not well understood what specific features of speech they use to do this. This paper introduces a data-driven framework to identify the time-frequency locations of these features. Using the same speech utterance mixed with many different noise instances, the framework is able to compute the importance of each time-frequency point in the utterance to its intelligibility. The mixtures have approximately the same global signal-to-noise ratio at each frequency, but very different recognition rates. The difference between these intelligible vs unintelligible mixtures is the alignment between the speech and spectro-temporally modulated noise, providing different combinations of "glimpses" of speech in each mixture. The current results reveal the locations of these important noise-robust phonetic features in a restricted set of syllables. Classification models trained to predict whether individual mixtures are intelligible based on the location of these glimpses can generalize to new conditions, successfully predicting the intelligibility of novel mixtures. They are able to generalize to novel noise instances, novel productions of the same word by the same talker, novel utterances of the same word spoken by different talkers, and, to some extent, novel consonants.

© *2016 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4964102]

[JFL] Pages: 2542–2553

I. INTRODUCTION

Normal-hearing listeners are remarkably good at understanding speech in noisy environments, much better than hearing-impaired listeners (e.g., Festen and Plomp, 1990; Alcántara *et al.*, 2003) and automatic speech recognition systems (e.g., Scharenborg, 2007). A better understanding of the robustness of normal hearing and an ability to reproduce it in machine listeners would likely enable improvements in theory as well as hearing aids and conversational interfaces. One theory of the mechanism underlying this process hypothesizes that listeners detect relatively clean "glimpses" of speech in the acoustic signal and assemble them into a percept (Cooke, 2006; Brungart *et al.*, 2006; Li and Loizou, 2007; Apoux and Healy, 2009). The current study is designed to reveal the locations of the glimpses that are most useful for correctly identifying particular utterances in noise, yielding a determination of "where" in the speech signal listeners find noise-robust phonetic information.

The techniques developed in this paper characterize the importance of individual time-frequency (T-F) points of a particular speech utterance by measuring its intelligibility when mixed with many different instances of a special "bubble" noise process. Auditory bubbles are designed to provide glimpses of the clean speech and to allow the measurement of importance of different glimpses of the same utterance. T-F points that are frequently audible in correctly identified mixtures and frequently inaudible in incorrectly identified mixtures are likely to be important for understanding that utterance in general. Because it is data-driven, results from this procedure can be compared across various conditions to compare listener strategies.

Two analyses are introduced to characterize these relationships, a correlational analysis and a predictive analysis. First, the correlational analysis identifies individual T-F points where audibility is correlated with overall intelligibility of the target word, or conversely, where noise is most intrusive or disruptive. This technique tends to identify a small number of such T-F points arranged in compact groups. Second, the predictive analysis uses information at each T-F point in a speech+noise mixture to predict whether that mixture will be intelligible or not. The goal is an ability to generalize to new mixtures, predicting better than chance the intelligibility of mixtures involving both new noise instances and new utterances. Figure 1 shows an overview of the bubble technique.

This work is inspired by methods from several fields. Healy *et al.* (2013) measured the band importance function for different speech corpora, and found that these functions were very consistent across listeners, but differed depending on the particular word/sentence material employed. However, traditional examinations of speech-band importance like ANSI (1997) and Healy *et al.* (2013) typically consider only differences across frequency bands and have

[a] Portions of this work were presented at the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, the 2014 ISCA Interspeech conference, and the 169th meeting of the Acoustical Society of America.

[b] Current address: Department of Computer and Information Science, Brooklyn College, CUNY, Brooklyn, NY 11210, USA. Electronic mail: mim@sci.brooklyn.cuny.edu

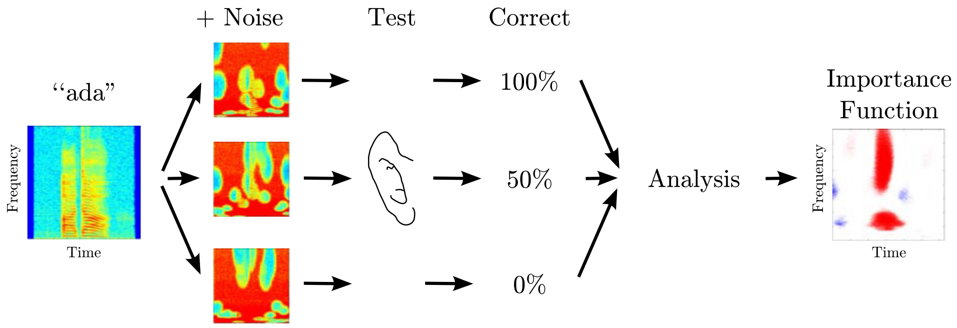


FIG. 1. (Color online) Overview of the proposed time-frequency importance function technique. The intelligibility is measured of an utterance mixed with many instances of noise with randomly placed "bubbles" excised from it. Correlation of the audibility of each point in the spectrogram with intelligibility across mixtures estimates the importance of each spectrogram point to the utterance's intelligibility in noise.

generally neglected the temporal aspect of these patterns. Ma *et al.* (2009) showed that *ad hoc* time-frequency weighting functions can improve the performance of objective predictors of speech intelligibility and Yu *et al.* (2014) showed that such weighting functions based on course groupings of speech and noise energy were similarly helpful. The data-driven importance values derived here should improve these predictions even further. Li *et al.* (2010) adopted the idea of measuring the intelligibility of the same utterance under a variety of modifications, including truncation in time and frequency and the addition of uniform noise. Because this technique involves truncation in time, it can only be applied to initial and final phonemes of utterances. In contrast, the currently proposed technique can be applied to phonemes in any position in a word, even in the context of running sentences.

Cooke (2009) found that certain combinations of speech and noise are recognized by multiple human listeners consistently as the same incorrect word, but that these misrecognitions were sensitive to the exact alignment of the speech and noise samples in time, fundamental frequency, and signal-to-noise ratio, suggesting that the localized time-frequency alignment of speech and noise can have large effects on intelligibility. The current "auditory bubbles" listening test methodology is based on the visual bubbles test (Gosselin and Schyns, 2001), which uses a visual discrimination task to identify the regions of images important for viewers to identify expressivity, gender, and identity.[1] The current predictive analysis is based on work such as Cox and Savoy (2003), in which classifiers are trained to predict the object classes seen by subjects from fMRIs of their brains obtained during the task.

The current study extends to time-frequency importance functions (TFIFs) methods that have been used in speech perception research to measure the importance of frequency bands averaged across many utterances (Doherty and Turner, 1996; Turner *et al.*, 1998; Apoux and Bacon, 2004; Calandruccio and Doherty, 2007; Apoux and Healy, 2012). These studies have been valuable in identifying the importance of various frequency bands of speech, averaged over time. Varnet *et al.* (2013) take this approach further by identifying time-frequency importance in the task of discriminating /b/ from /d/. Their results showed that the transition of the second formant was key for performing this task, in agreement with traditional views of speech cues, and furthermore identified that this estimation was performed relative to the same formant in the previous syllable. Their use of white Gaussian noise as the corruption signal, however, required an order of magnitude more trials than the technique proposed here, which uses noise with larger time-frequency modulations.

The purposes of the current study are to establish the bubbles technique for measuring the TFIF, to examine time-frequency regions of importance, both positive and negative—those that support the identification of specific consonant sounds and those that are potentially misleading to accurate identification, and to determine if a machine classifier can predict human performance based on the specific information spared disruption by bubble noise.

II. EXPERIMENTS

A. Method

1. Subjects

Subjects were 13 volunteers having normal hearing as defined by audiometric thresholds on day of test $\leq$20 dB hearing level (HL) at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010). They were females aged 18–22 years and participated for extra course credit.

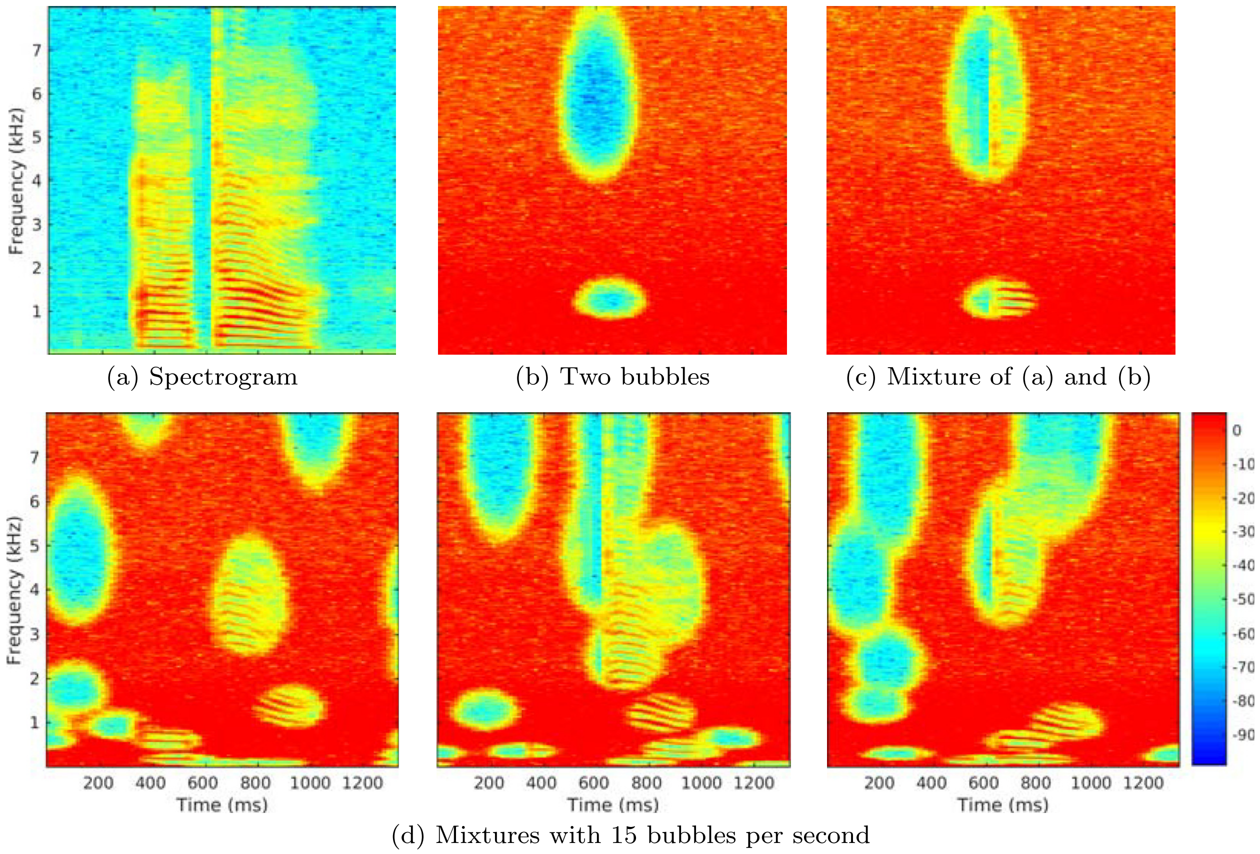
2. Stimuli

The speech material was selected from the corpus described by Shannon *et al.* (1999) and consisted of several pronunciations of six vowel-consonant-vowel (VCV) nonsense words. The nonsense words were of the form /aCa/: /ɑtʃɑ/, /ɑdʒɑ/, /ɑdɑ/, /ɑtɑ/, /ɑfɑ/, /ɑvɑ/. This limited stimulus set was selected for the current initial test of the bubble technique, to allow focus on optimizing the method and to ensure interpretable patterns of results. Three productions of each word came from a single female talker (number W3). The longest- and shortest-duration versions of each utterance were selected from approximately 10 versions along with one of an intermediate duration, designated "v1," "v2," and "v3" from shortest to longest. Three more productions of the same words came from three different talkers, numbers W2, W4, and W5. These talkers were selected because their recordings were of the highest apparent quality and they showed large variation in speaking style. Talkers of the same gender were selected so that they had similar voice pitches and formant positions. Female talkers were selected because they had fewer pronunciation errors than the male talkers. The stimuli were all 2.2 s in duration including surrounding silence. The plots show the central 1.2 s, which is sufficient to include all of the speech content. The various productions were roughly centered within the stimuli, but were not temporally aligned in any way beyond that (except during the machine learning analysis methodology, as described in Sec. II B 3). The signals were sampled at 44.1 kHz and 16 bits.

Each utterance was mixed with multiple instances of "bubble" noise. This noise was designed to provide glimpses of the speech only in specific time-frequency bubbles. This noise began as speech-shaped noise with an SNR of –27.5 dB, sufficient to make the speech completely unintelligible. The noise was then attenuated in "bubbles" that were jointly parabolic in time and ERB$_\mathrm{N}$-scale frequency (Glasberg and Moore, 1990) with a maximum suppression of 80 dB. The center points of the bubbles were selected uniformly at random locations in time and in ERB$_\mathrm{N}$-scale frequency, except that they were excluded from a 2–ERB$_\mathrm{N}$ buffer at the bottom and top of the frequency scale to avoid edge effects (no frequency limits were imposed outside of Nyquist). Mathematically, the attenuation applied to the speech-shaped noise, $M(f, t)$, is

$$B(f,t) = \sum_{i=1}^{I} \exp\left\{ -\frac{(t-t_i)^2}{\sigma_t^2} - \frac{(E(f)-E(f_i))^2}{\sigma_f^2} \right\},$$
$$M(f,t) = \min\left(1, \frac{10^{-80/20}}{B(f,t)}\right), \qquad (1)$$

where $E(f) = 21.4\log_{10}(0.00437f + 1)$ converts frequencies in Hz to ERB$_\mathrm{N}$, and $\{(f_i, t_i)\}_{i=1}^{I}$ are the randomly selected centers of the I bubbles. The scale parameters σ_t and σ_f were set such that the bubbles were fixed in size to have a half-amplitude "width" of 90 ms at their widest and a half-amplitude "height" of 1 ERB$_\mathrm{N}$ at their tallest, the smallest values that would avoid introducing audible artifacts. For the full 80-dB dynamic range, this corresponds to 350 ms wide at their widest and 7 ERB$_\mathrm{N}$ high at their highest. Future experiments could explore the use of lower maximum suppression values along with smaller bubbles, which should increase the resolution of the method, but might require more mixtures per token. The number of bubbles was set such that listeners could correctly identify approximately 50% of the mixtures from the six-alternative forced choice. Pilot experiments showed that approximately 15 bubbles per second achieved this level of identification, which led to a final overall SNR of −24 dB. Figure 2 displays spectrogram images of this bubble noise. Figure 2(b) shows a noise having only two bubbles and Fig. 2(c) shows the utterance in



(a) Spectrogram

(b) Two bubbles

(c) Mixture of (a) and (b)

(d) Mixtures with 15 bubbles per second

FIG. 2. (Color online) Example bubble-noise instances and mixtures with the word /ɑdɑ/. (a) Spectrogram of clean utterance. (b) An example instance of bubble noise with two bubbles only. (c) The mixture of the utterance in (a) with the noise in (b). (d) Three mixtures of the utterance with bubble noise having 15 bubbles per second.

(a) mixed with this noise. In Fig. 2(d), the same utterance is mixed with three instances of bubble noise having 15 bubbles per second and randomly determined bubble locations.

3. Procedure

Subjects were seated in a double-walled IAC sound booth in front of a computer running a custom MATLAB presentation interface. Sounds were presented diotically via Echo D/A converters (Santa Barbara, CA) and Sennheiser HD280 PRO headphones (Wedemark, Germany). Sound presentation levels were calibrated via a Larson Davis (Depew, NY) sound level meter and coupler so that mixtures were presented at 75 dBA. One mixture at a time was selected for presentation at random from those assigned to the listener. The listener then selected the word that they heard from the closed set of six using a textual MATLAB interface. Testing began with two 5-min training periods. The first used the clean utterances and the second used bubble-noise utterances. Feedback was provided during training, but not during the formal testing.

a. Experiment 1. Intra-subject consistency was measured through repeated presentation of mixtures. Five subjects participated in this experiment. Five specific speech+noise mixtures for each of the six medium-rate words spoken by talker W3 were presented 10 times to each listener. Thus, each listener performed 300 labelings for this experiment. The proportion of those 10 presentations in which each mixture was correctly identified was then computed, leading to 30 proportions per listener.

b. Experiment 2. Inter-subject consistency was measured through repeated presentation of different mixtures. This experiment was performed by the same five subjects as experiment (exp.) 1 using the same six clean utterances. Two hundred mixtures of each of the six utterances were generated. Each of these 1200 mixtures used a unique bubble noise. Every mixture was presented to each listener once and the agreement between listeners on each mixture was quantified using Cohen's κ (Cohen, 1960).

c. Experiments 3a and 3b. Eight listeners who were not involved in exps. 1 and 2 participated in an experiment to assess importance. Four listeners participated in exp. 3a involving the 18 utterances from talker W3 described in Sec. II A 2 (six words × three utterances). The other four listeners participated in exp. 3b involving the 18 utterances from talkers W2, W4, and W5 (six words × three utterances). Results from these experiments were analyzed together. Each listener was assigned 50 unique mixtures of each of their 18 utterances. Testing to identify all 900 mixtures (18 utterances × 50 mixtures) took approximately one hour. Thus, together, exps. 3a and 3b included 7200 unique mixtures, distinct from those used in exps. 1 and 2. Listeners responded to each mixture by selecting the word heard from the six choices. Because each mixture was only heard a single time, it was considered to be intelligible if it was correctly identified and unintelligible if not.

B. Analytical approaches for importance assessment

The extraction of importance information from the current listening tests involved a correlational approach to identify compact important regions in the spectrogram of utterances and a predictive approach to predict whether novel mixtures will be intelligible to human listeners based on the particular arrangement of speech and noise in time and frequency.

1. Correlational: Point-biserial correlation

The first analysis involved an examination of the correlation between the audibility at each individual T-F point and the correct identification of the word in the corresponding mixture. Audibility was quantified as the difference between the level of the original speech-shaped noise and that of the bubble noise, i.e., the depth of the "bubbles" in the noise at each T-F point. Following Calandruccio and Doherty (2007), the point-biserial correlation was used for this calculation, which computes the correlation between a dichotomous variable (correct identification of mixture) with a continuous variable (audibility at a given T-F point). The significance of this correlation can also be tested using a one-way analysis of variance with two levels, with p-value denoted $p(f,t)$. The degree to which the audibility of a particular T-F point is correlated with correct identification of the word should indicate the importance of that T-F point to the intelligibility of that word. In contrast, points where audibility is not significantly correlated with correct identification are likely not as important to its intelligibility.

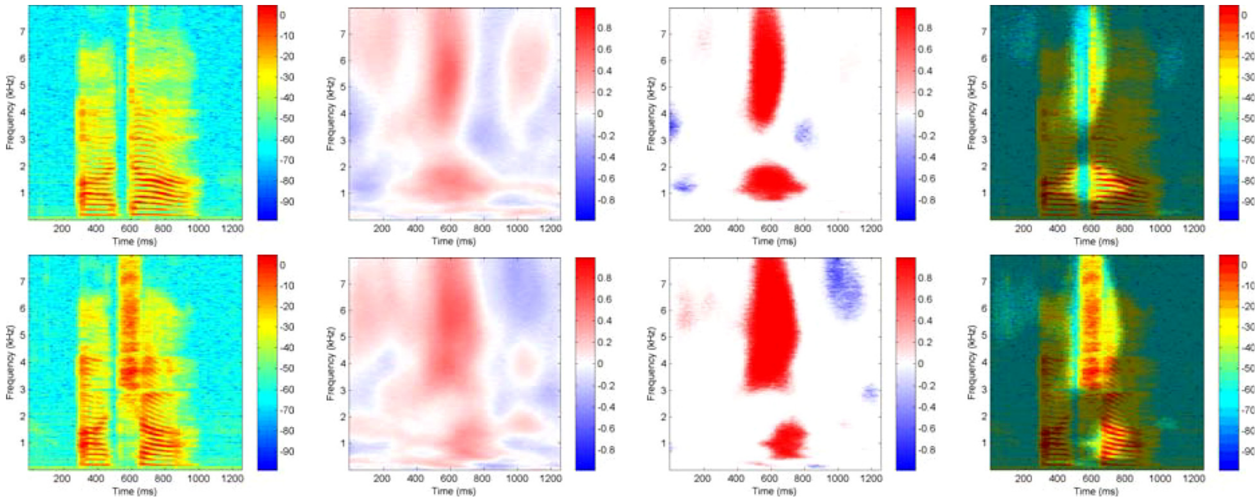
Figure 3 shows several visualizations of the correlational analysis for one utterance of the nonsense words /ɑdɑ/ and /ɑtʃɑ/. Figure 3(a) shows the spectrogram of the clean utterance. Figure 3(b) shows the correlation between audibility at each T-F point across mixtures involving this utterance and the intelligibility of each mixture, with positive correlations in red and negative correlations in blue. Figure 3(c) shows the quantity $M_\chi(f,t) = \exp[-p(f,t)/0.05]$ a visualization of the significance of this correlation at each T-F point. And Fig. 3(d) reveals the original clean spectrogram through the portions of $M_\chi(f,t)$ that show positive correlations between audibility and intelligibility.

2. Predictive machine learning

The second method employed to compute an intelligibility map used a linear classifier, in this case a support vector machine (SVM). This machine learning method is predictive because, in contrast to the correlational method, it allows the quality of the fit to be measured on data not used in the training process via the prediction accuracy of the model.

All of the mixtures involving a given clean recording constituted a single learning problem. The features used were $G_m(f,t)$, the amount that the speech-shaped noise had been suppressed by its bubbles as a function of frequency and time in the mth mixture. The machine learning task is to predict whether the mth mixture was intelligible, denoted y_m. Because all of the features considered in a single problem corresponded to the same clean recording, these features implicitly represented the speech and did not need to explicitly represent it.

Because of the large number of dimensions of the $G_m(f,t)$ (513 frequencies × 64 frames = 32 832 dimensions), the first stage of analysis was a dimensionality reduction using principal components analysis (PCA), which resulted in 5 to 120 dimensions. Computing PCA on the features directly



(a) Spectrogram (b) Correlation (c) Significance (d) Visualization

FIG. 3. (Color online) Example of correlational analysis for /ɑdɑ/ (top panels) and /ɑtʃɑ/ (bottom panels) from talker W3. (a) Original spectrogram. (b) Correlation between noise level at each point and consensus intelligibility of the mixture. (c) Significance of this correlation, with significant positive correlations in red and negative in blue. (d) Positive significance from (c) translated to transparency in a mask and overlaid on original spectrogram.

gave too much weight to the high-frequency bubbles, because the same number of ERB_N's contained a larger bandwidth in Hz and so many more short-time Fourier transform (STFT) frequency channels in the high-frequency bubbles relative to the low-frequency bubbles. The features were thus reweighted before performing PCA to counteract this effect. The weight used was the cube root of the incremental ERB_N frequency change between adjacent STFT frequency channels. This frequency weighting method is similar to frequency warping methods like those used in computing mel frequency cepstral coefficients (Davis and Mermelstein, 1980), but without the loss of information caused by combining many high-frequency channels together before the analysis.

Experiments to reconstruct individual bubble-noise instances from this PCA approximation showed that two PCA dimensions per bubble in the instance led to accurate reconstructions. This makes sense, as since bubbles are a fixed size, each bubble effectively encodes two independent dimensions of information: the frequency and time of its center. Thus bubble noise instances occupy a very low-dimensional, relatively linear subspace of the much higher dimensional spectrogram space.

The SVM classifier used here is known to be sensitive to an imbalanced number of positive and negative examples in its training data (Rehan Akbani *et al.*, 2004), so training points from the over-represented class were discarded to achieve balance, as is typical. Thus, if listeners achieve accuracy greater than 50%, the predictive analysis will only be able to utilize a number of training examples equal to twice the number of incorrectly identified mixtures (and vice versa). Note that in contrast to the listeners' task of selecting the correct word of six, which has a chance level of 16.7%, the classifier is predicting whether a particular utterance was correctly identified or not, which has a chance level of 50%, because of our tuning of the number of bubbles per second.

Nested cross-validation was used to select the PCA dimensionality on the training data of each cross-validation fold that led to the best classification performance. In particular, the data were partitioned into five approximately equal collections of examples. Models using different parameters (dimensionality in this case) were trained on three of these. The model that performed best on the fourth, the development set, was selected to be officially evaluated on the fifth, the test set. This procedure was then repeated five times, using each set as the test set in turn, and averaging the results on all of the test sets together. In this way, model parameters can be selected in a data-driven way that is still independent of the test set, giving a fair evaluation. The dimensionality selected was generally between 12 and 31, with 31 being the most common by a small margin. If the linear classifier is

$$\hat{y}_m = b + \sum_k w_k \sum_{f,t} B_k(f,t) G_m(f,t), \qquad (2)$$

where $B_k(f,t)$ is the kth PCA basis, then the corresponding intelligibility map is $M_s(f,t) = \sum_k w_k B_k(f,t)$.

3. Alignment: Dynamic time warping

Some processing was necessary in order to permit generalization between different utterances. This is because the features used in the various analyses only represent the clean speech implicitly. In cross-utterance experiments, one utterance was selected as the reference and the others were continuously warped in time to match it. This is true both for experiments across different productions of the same word and for experiments across different words. Specifically, a time warp for a source utterance was computed using the MATLAB code of Ellis (2003) to minimize the sum of squared errors between its mel frequency cepstral coefficients and those of the target utterance, with no penalty for insertions or deletions. This warp was then applied to the features of the source utterance's mixtures before performing the predictive analysis. In general, additional transformations could be used, including the alignment of pitch and vocal tract length across utterances, but such transformations were not used in the current studies.

III. RESULTS

A. Experiment 1: Intra-subject consistency

Figure 4 shows the intra-subject consistency results from exp. 1. Recall that this experiment measured the proportion of 10 presentations of the same mixture that each listener was able to correctly identify. These 150 proportions (30 proportions × 5 listeners) are displayed in the histogram in Fig. 4(a). Proportions at or below chance are grouped together into the lowest bin. The U-shaped curve indicates that most mixtures were either highly intelligible or unintelligible, and that listeners were consistent in this ability or inability to identify the word.

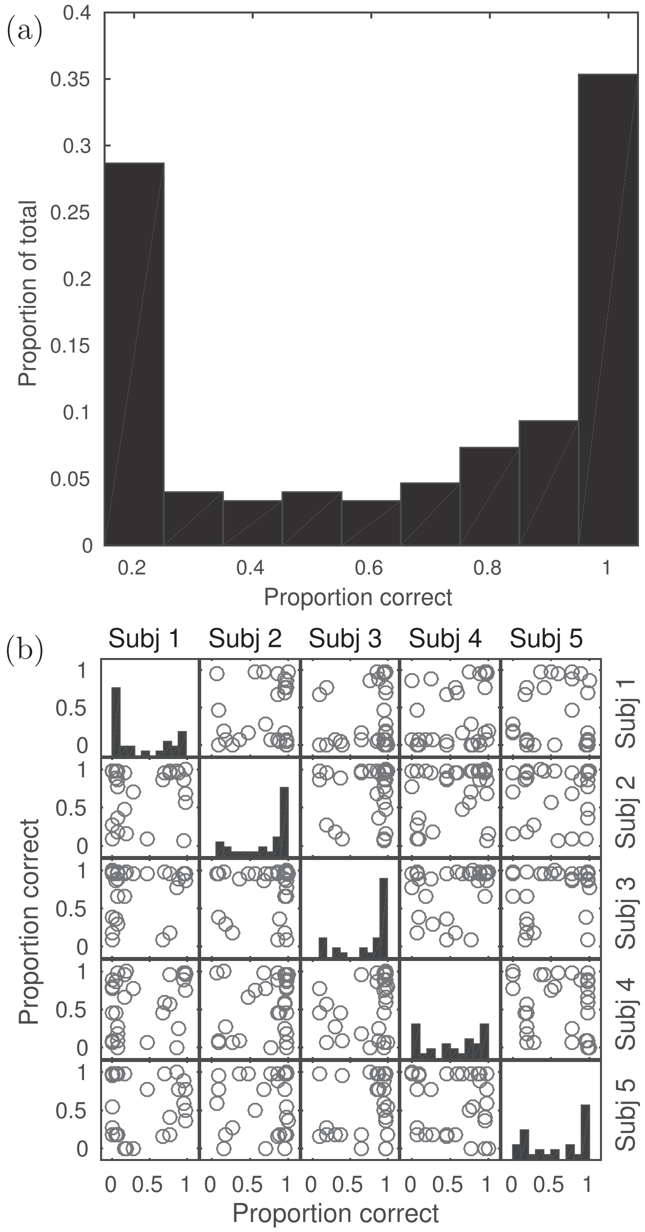


FIG. 4. Results of exp. 1: intra-subject consistency. (a) Intra-subject consistency. The U-shaped curves show that most mixtures were either highly intelligible, or unintelligible. (b) Inter-subject consistency of proportion correct for each of the 30 repeated mixtures, showing a good deal of agreement between subjects 2–5, especially on correct identifications.

Figure 4(b) shows the same data broken down by listener. The plots on the diagonal of the matrix show the same kind of histogram as in Fig. 4(a), but for each listener individually. They again show a good deal of intra-subject consistency. Off of the main diagonal, the scatter plots address inter-subject consistency. They show the proportion correct for one listener versus the proportion correct for another listener, on the same mixtures. These plots show modest agreement on intelligibility between subjects, especially on correct identifications, i.e., mixtures that were intelligible to one listener tended to be intelligible to the others. Note that the y-axis proportions correspond to those of the scatter plots and not the histograms, although all of the histograms use a consistent scale.

B. Experiment 2: Inter-subject consistency

Figure 5 displays the inter-subject consistency results from exp. 2. It is similar to Fig. 4(a), except that it shows the proportion of listeners who were able to correctly identify each of the 1200 mixtures in exp. 2. Similarly to Fig. 4(a), the slight U shape to the curve indicates that mixtures tended to be either intelligible (most cases) or unintelligible.

Analysis of the results from exp. 2 using Cohen's κ (Cohen, 1960) are shown in Table I. Cohen's κ measures agreement between two subjects normalized for chance levels of agreement due to the marginal distribution of each response. We measure κ for the six-way responses of the listeners (i.e., whether two listeners agree on the selected word for a given mixture). In general, κ values between 40% and 60% are considered to be moderate and those between 60% and 80% are considered substantial (Landis and Koch, 1977). Table I shows the value of κ for each pair of listeners on the 1200 mixtures in exp. 2. These results again show a large amount of agreement between subjects, although subjects 2–5 tend to agree more with each other than with subject 1.

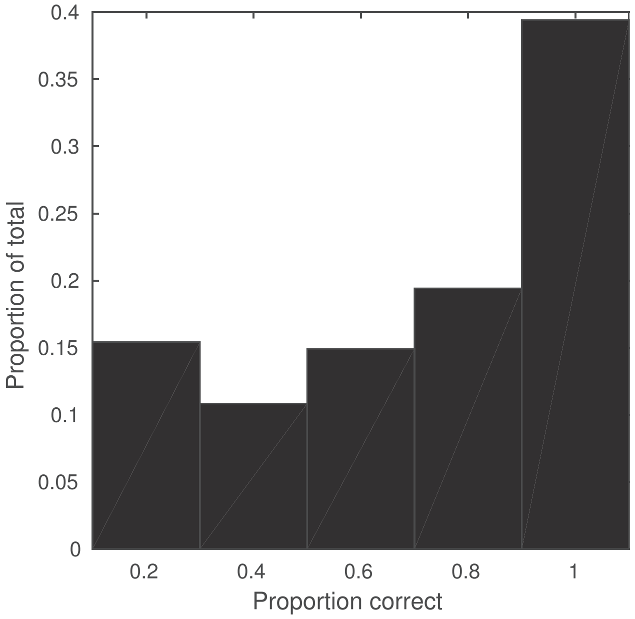


FIG. 5. Results of exp. 2: inter-subject consistency, measured as the proportion of subjects who correctly identified each of the 1200 mixtures in exp. 2. The slight U-shape to the curve indicates that mixtures tended to be either intelligible (most cases) or unintelligible.

TABLE I. Cohen's κ (as a percentage) measuring consistency between pairs of subjects (subj.) on responses (six-way) on the 1200 mixtures from exp. 2. These results again show a large amount of agreement between subjects, especially among subjects 2–5, with subject 1 only showing moderate agreement with the others.

Subj.	1	2	3	4	5
1		46.3	52.6	50.9	51.1
2	46.3		58.3	56.1	56.8
3	52.6	58.3		64.0	68.8
4	50.9	56.1	64.0		61.2
5	51.1	56.8	68.8	61.2	

Figure 6 shows time-frequency importance functions (described in Sec. II B 1) of the word /ɑdɑ/ computed for each of the five listeners using the same exp. 2 mixtures used to construct Fig. 5. The sixth TFIF is derived from the consensus intelligibility results. In this consensus, a mixture was considered to be intelligible if all five listeners correctly identified it and was considered unintelligible if at least two listeners could not correctly identify it. Mixtures that were correctly identified by all but one listener were ignored for the purposes of the consensus analysis. This grouping of "votes" resulted in approximately equal numbers of mixtures categorized as intelligible and unintelligible. These plots show that the TFIFs derived from the responses of each listener are quite similar.

Table II shows the percentage of mixtures correctly identified for each clean utterance in exp. 2. Note that many of these percentages are above the target value of 50%. As discussed in Sec. II B 2, the effect of this is to reduce the number of mixtures per utterance that can be utilized in some of the analyses. There is some variation in intelligibility across different listeners, especially listener 1, which might explain some of the inter-subject differences identified by Cohen's κ and shown in Table I.

C. Experiment 3: Assessing importance

The importance assessment of exps. 3a and 3b employed more clean utterances (six productions of each word) and presented each mixture to only a single listener. Table III shows the percentage of mixtures correctly identified for each clean utterance, averaged across subjects. Note that, as in Table II, many of these percentages are above the target value of 50%. Note also that there is a large degree of variation in intelligibility across different utterances, and across talkers. Talker W2 appears to be less intelligible than the others at this particular number of bubbles per second.

TABLE II. Percent of mixtures correctly identified by listeners in exp. 2—inter-subject consistency.

Subject	/ɑtʃɑ/	/ɑdʒɑ/	/ɑdɑ/	/ɑtɑ/	/ɑfɑ/	/ɑvɑ/	Avg.
1	49.5	54.0	67.5	62.5	64.5	65.0	60.5
2	67.0	66.0	59.5	65.0	65.5	69.0	65.3
3	79.5	70.5	79.5	79.5	84.5	74.5	78.0
4	68.5	65.5	65.0	77.5	77.0	71.5	70.8
5	72.0	65.5	64.5	75.0	90.5	83.0	75.1
Consensus	67.3	64.3	67.2	71.9	76.4	72.6	69.9

1. Correlational analysis

Figure 7 shows the time-frequency importance functions for all of the utterances used in exps. 3a and 3b. Each importance function was derived from 200 mixtures involving each word (4 listeners × 50 mixtures each). Of these 36 TFIFs, 18 were derived from the same-talker utterances in exps. 3a and 18 from the different-talker utterances in exp. 3b. Each row shows a different word and each column a different utterance. The spectrogram of the original utterance is shown in color, with the importance represented as the level of the "value" in the hue-saturation-value color model, i.e., the colors are darkened in unimportant regions.

2. Predictive analysis

The first analysis involved a baseline in which the machine-learning classifiers were trained and tested on the same utterance (a given production of a given word by a given talker) and had to generalize to only new bubble-noise mixtures. Table IV shows the accuracy of the classifiers when trained and tested on mixtures from exp. 3 involving the same clean speech utterance, using fivefold cross-validation. The classifiers were trained on 80% of the mixtures and tested on the remaining 20%, with the training and testing divisions rotated through the five possibilities and the accuracies averaged. Approximately half of the accuracies for individual mixtures in Table IV are significantly above chance levels of 50% at a 0.05 level according to a one-sided binomial test. The averages across talkers, however, are significantly above chance for each word according to the same test. This analysis shows the accuracy that classifiers can achieve when required to generalize only across noise instances and not across speech utterances.

Table V shows the cross-utterance classification accuracy of the classifiers. In contrast to Table IV, the accuracy of these classifiers is measured when predicting the intelligibility of novel mixtures that include both noise instances and

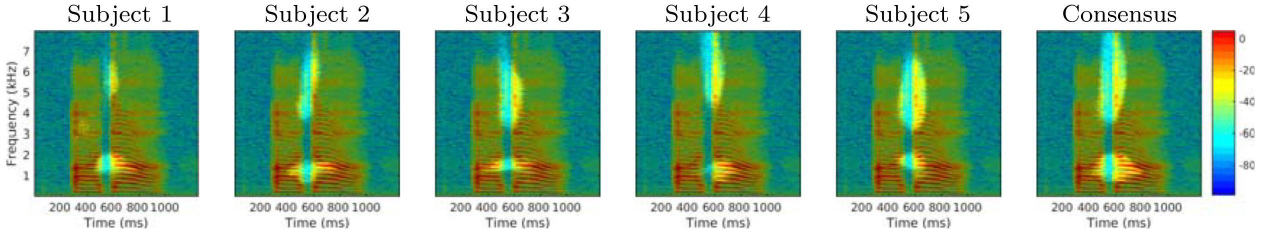


FIG. 6. (Color online) Time-frequency importance functions [as in Fig. 3(d)] of /ɑdɑ/ computed for five different listeners, each hearing the same mixtures in exp. 2, and computed from the consensus intelligibility estimate.

TABLE III. Percent of mixtures correctly identified by listeners in exps. 3a and 3b—Importance assessment.

Talker	/ɑtʃɑ/	/ɑdʒɑ/	/ɑdɑ/	/ɑtɑ/	/ɑfɑ/	/ɑvɑ/	Avg.
W3 v1	72.0	70.5	60.5	75.0	60.0	70.5	68.1
W3 v2	72.0	69.5	64.5	72.0	64.5	76.5	69.8
W3 v3	74.0	64.5	67.5	50.0	57.0	69.5	63.8
W2	32.0	49.5	52.5	58.5	32.0	75.0	49.9
W4	75.0	54.5	75.5	77.0	64.0	56.0	67.0
W5	52.5	47.0	62.5	73.5	51.0	67.5	59.0
Avg.	62.9	59.3	63.8	67.7	54.8	69.2	62.9

utterances that were not used in the training of the classifier, thus measuring its ability to generalize to both new noise instances and new utterances. In each case, a model was tested on the 200 mixtures involving a single utterance from exp. 3 after being trained on the mixtures involving the other utterances of the same nonsense word from exp. 3. This procedure was repeated, rotating through each of the utterances for testing, with the accuracies averaged together. The same-talker conditions used the utterances and results from exp. 3a, in which only talker W3 was employed. The different-talker conditions used the utterances and results from exp. 3b (talkers W2, W4, W5) plus the "v1" utterances from exp. 3a, which represent a fourth unique talker. Thus, in the same-talker condition, the classifiers were trained on the 400 mixtures involving two utterances of each word from talker W3 and tested on the 200 mixtures involving the third utterance from that talker. In the different-talker condition, the classifiers were trained on the 600 mixtures involving three productions of each word, each from a different talker, and tested on the 200 mixtures involving the fourth production from the fourth talker. Table V shows these results with and without the time alignment described in Sec. II B 3 performed to W3 v3 for the single-talker condition and W4 for the multiple-talker condition. It shows that these classifiers are able to generalize across different utterances of the same word spoken by both the same and different talkers, and that they are better able to do so when the utterances are time-aligned to a reference. Subsequent cross-utterance results will therefore only be reported with the use of time warping.

Careful examination of Tables IV versus V indicates that the accuracy of the classifiers was lower when they were required to generalize only across noise instances than when they were required to generalize across noise instances and utterances. This is even true for five of the six words in the

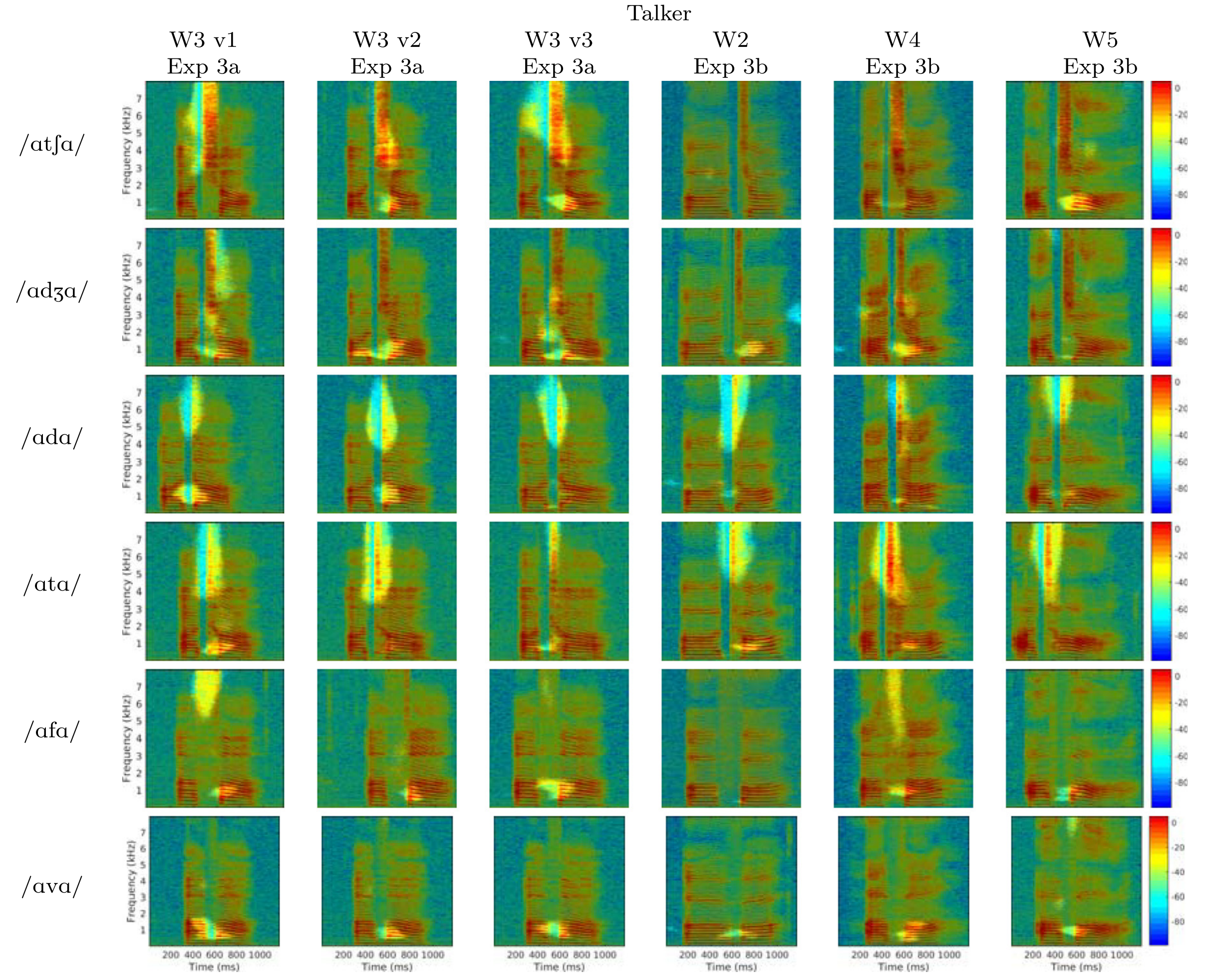


FIG. 7. (Color online) Time-frequency importance functions derived from exps. 3a and 3b data, overlaid on spectrograms of each of the 36 utterances. Utterances from talker W3 v2 were used in exps. 1 and 2.

TABLE IV. Machine learning classification accuracy (using 5-fold cross-validation) within individual utterances. Data were from exp. 3. Bold entries are significantly above chance performance (50%) at a 0.05 level according to a one-sided binomial test.

Talker	/ɑtʃɑ/	/ɑdʒɑ/	/ɑdɑ/	/ɑtɑ/	/ɑfɑ/	/ɑvɑ/	Avg.
W3 v1	54.5	**65.3**	**58.9**	49.0	53.8	54.2	**55.9**
W3 v2	58.0	**68.0**	**65.5**	**62.5**	58.5	59.6	**62.0**
W3 v3	**64.4**	**66.2**	58.5	**66.5**	**64.0**	56.6	**62.7**
W2	57.8	49.5	**73.2**	**62.7**	52.3	55.0	**58.4**
W4	60.0	**61.0**	54.1	**71.7**	57.6	**58.5**	**60.5**
W5	**59.5**	**61.2**	56.7	**64.2**	52.6	53.8	**58.0**
Avg.	**59.0**	**61.9**	**61.1**	**62.8**	**56.4**	**56.3**	**59.6**

different-talker conditions. This apparent anomaly may be related to the different amounts of data available to train the classifiers. Table VI shows the accuracy of cross-utterance classifiers trained on the same number of examples as the within-utterance classifiers as in Table IV. The number of training points, averaged across utterances, is shown in the N_{tr} column. Note that N_{tr} for the within-utterance rows is lower than 160 examples because of the balancing of the number of examples between the positive and negative classes of the training data. These results show that the increase in accuracy of the cross-utterance classifiers over the within-utterance classifiers is essentially eliminated when the number of training examples is reduced to that of the within-utterance classifiers.

Table VII compares the ability of these classifiers to generalize across different versions of the same word with their ability to generalize across different words. It includes results for both the same talker and different talkers. Significance is tested using a one-sided binomial test comparing against a baseline of 50% accuracy with a significance level of 0.05. In the different-word condition, each of the classifiers trained in the same-word condition (in the same way as for Tables V and VI) is tested on three randomly selected utterances of different words. These random words come from the same talker in the same-talker condition and from different talkers in the different-talker condition. The bottom half of the table compares various pairs of rows from the top half, with significant differences in bold. The significance test used in this case is a two-sided two-proportion z-test with a significance level of 0.05. These results show that cross-utterance classifiers are able to correctly predict intelligibility of individual unseen mixtures at better than chance levels for all conditions, but especially when training and testing on mixtures from the same word.

TABLE V. Cross-utterance classification accuracy on mixtures involving a novel utterance. Data were from exp. 3. Same-talker models were trained on two utterances, different-talker models on three. Results are shown with (+) and without (−) aligning the clean utterances to a reference. All results are significantly better than chance at a 0.05 level according to a one-sided binomial test.

Talker	Warp	/ɑtʃɑ/	/ɑdʒɑ/	/ɑdɑ/	/ɑtɑ/	/ɑfɑ/	/ɑvɑ/	Avg.
Same	+	**65.2**	**62.7**	**74.1**	**74.4**	**67.0**	**63.5**	**67.8**
Same	−	**63.6**	**63.9**	**63.4**	**74.0**	**57.2**	**64.5**	**64.4**
Diff.	+	**62.3**	**58.7**	**69.3**	**73.9**	**59.9**	**61.6**	**64.3**
Diff.	−	**55.7**	**58.9**	**66.8**	**63.1**	**61.6**	**59.9**	**61.0**

TABLE VI. Accuracy of cross-utterance classification and cross-validation within utterances showing the effect of increasing the amount of training data. Data were from exp. 3. Abbreviations: Talker (T), utterance (U), same (S), and different (D). In the case of the different-utterance condition, all utterances were still of the same word. Bold entries are significantly better than chance at a 0.05 level according to a one-sided binomial test.

T	U	N_{tr}	/ɑtʃɑ/	/ɑdʒɑ/	/ɑdɑ/	/ɑtɑ/	/ɑfɑ/	/ɑvɑ/	Avg.
S	S	104.9	**59.0**	**66.5**	**60.9**	**59.3**	**58.7**	**56.8**	**60.2**
S	D	105.1	**57.6**	**62.0**	**69.5**	**63.7**	**56.5**	**60.0**	**61.5**
S	D	262.2	**65.2**	**62.7**	**74.1**	**74.4**	**67.0**	**63.5**	**67.8**
D	S	114.3	**57.9**	**59.2**	**60.7**	**61.9**	**54.1**	**55.4**	**58.2**
D	D	114.1	50.7	49.6	**54.8**	**59.0**	**55.1**	53.9	**53.8**
D	D	464.7	**62.3**	**58.7**	**69.3**	**73.9**	**59.9**	**61.6**	**64.3**

IV. DISCUSSION

A. Listener consistency

The U-shaped histogram in Fig. 4(a) indicates that most speech+noise mixtures were either highly intelligible, or unintelligible, and that there were not many that were in-between. Thus, on repeated presentations, listeners tended to repeatedly identify a given mixture correctly or incorrectly. This supports the idea that certain combinations of glimpses allow a particular word to be correctly identified, whereas others do not. The plots along the diagonal of Fig. 4(b) confirm the results displayed in Fig. 4(a), by showing that mixtures tend to have high intelligibility or low intelligibility for each listener.

The off-diagonal plots in Fig. 4(b) show the relationship between the proportion correct of given mixtures when heard by different listeners. Each panel shows one pair of listeners. The clustering of points in the upper-right corner of each panel indicates that listeners tend to find the same mixtures intelligible, but seem to find different mixtures unintelligible. This could be a result of the slight skew toward correct identification.

Figure 5 shows that many of the mixtures can be correctly identified by all of the listeners, and mixtures incorrectly identified by many listeners are less frequent. This result agrees to some extent with the intra-subject consistency result in Fig. 4(a), but it is less definitively U-shaped.

TABLE VII. Cross-utterance accuracy for various combinations of same and different talker and test word, along with comparisons between them (Δ). Data were from exp. 3. Bold entries are statistically significant.

Word	Talker	/ɑtʃɑ/	/ɑdʒɑ/	/ɑdɑ/	/ɑtɑ/	/ɑfɑ/	/ɑvɑ/	Avg.
Same	Same	**65.2**	**62.7**	**74.1**	**74.4**	**67.0**	**63.5**	**67.8**
Same	Diff	**62.3**	**58.7**	**69.3**	**73.9**	**59.9**	**61.6**	**64.3**
Diff	Same	**62.2**	**59.8**	**61.8**	**57.5**	**61.9**	**56.8**	**60.0**
Diff	Diff	**64.0**	**62.0**	**61.4**	**58.0**	**59.9**	**59.4**	**60.8**
Δ	Same	3.0	2.8	**12.3**	**16.9**	5.2	**6.7**	**7.8**
Δ	Diff	−1.8	−3.3	**7.9**	**15.8**	0.0	2.2	**3.5**
Same	Δ	2.9	4.0	4.9	0.6	**7.1**	1.9	**3.6**
Diff	Δ	−1.8	−2.1	0.5	−0.5	2.0	−2.6	−0.8

It is likely that the relatively high overall recognition rate of 70%, as shown in Table II, skewed the U-shape toward more correct responses.

As can be seen from Table I, the κ values are between 46% and 69%, which further indicates that inter-subject agreement is generally high. Subject 1 has the lowest agreement with other subjects, especially subject 2, while subjects 3, 4, and 5 agree with each other substantially.

All listeners' TFIFs for the nonsense word /ɑdɑ/ in Fig. 6 show significant correlations between intelligibility and audibility in two distinct regions. One region is high in frequency, generally centered around 5–6 kHz, and corresponds to the closure and plosive-release burst of the consonant. The other region is lower in frequency, centered around 1.5 kHz, and corresponds to the transition of the first and second formants into and out of the consonant.

B. Correlational analysis

The maps in Fig. 7 display the cues used by listeners when extracting consonants from noise. They show many notable properties that agree with traditional theories of speech perception in many regards, while providing additional insight into the speech perception process in others. All utterances appear to have one or two main regions of importance, focused on similar information across utterances, but at different time-frequency regions because of idiosyncratic differences in the utterances. Specifically, there is a high-frequency region focused on frication and burst energy, and a low-frequency region focused on the first two formants and their transition before, during, and especially after the consonant. The low-frequency region seems to be present with some consistency, while the high-frequency region is less consistent, with different productions of the same word tending to be somewhat consistent in the presence or absence of the high-frequency region.

The high-frequency importance region is present in all productions of /ɑdɑ/ and /ɑtɑ/, in some productions of /ɑtʃɑ/, but in fewer productions of /ɑdʒɑ/, /ɑfɑ/, and /ɑvɑ/. This could reflect the listeners adopting a task-specific strategy for recognizing these words. The words /ɑfɑ/ and /ɑvɑ/ have relatively steady second-formant transitions after the consonant, while the other words have falling second-formant transitions, meaning that hearing a steady second-formant transition would be sufficient to distinguish these fricatives from the words with other manners of articulation. Such a hypothesis is in accord with the lack of importance found for the initial and final vowels, as they do not aid in discriminating these nonsense words.

The low-frequency regions are present for almost all of the words. They are only absent when there is little importance in an entire utterance, such as for /ɑtʃɑ/ and /ɑfɑ/ from talker W2, which is likely a result of too few examples of either intelligible or unintelligible mixtures leading to a lack of statistical power. In the case of those two words, Table III shows that listeners correctly identified only 32% of these mixtures. This is only slightly above the chance rate of 17%, meaning that these mixtures were overall very difficult to correctly identify using the same number of bubbles per second as the other utterances.

For utterances where 200 mixtures provided enough statistical power to create a reasonable importance map, the low-frequency importance regions almost always encompass the formant transitions from the consonant to the final vowel. For many of the voiced consonants, these regions also encompass the formants at the end of the initial vowel and span the low-energy gap in the consonant. Thus, attendance to these regions could indicate an assessment of voice onset time or, more specifically, the amount of time between the voicing offset of the initial vowel and onset of the final vowel. The location of these areas of importance suggests that consonant offsets are used by listeners more than consonant onsets when distinguishing VCV nonsense words. The fact that this low-frequency region was important even for consonants that are typically assumed to contain little phonetic information in this region (e.g., /f/) has potentially important implications for the human extraction of speech from noise. One implication of this finding is that listeners are attending not only to where information is, but also to where it might occur, in order to distinguish between possible utterances. If this listening strategy is extrapolated from the task employed currently to one involving the recognition of open-set utterances (i.e., sentences), it becomes clear that substantial areas of the spectrogram would have to be monitored.

Whereas the overlays in Fig. 7 only show the regions where audibility is positively correlated with intelligibility, there are some utterances for which audibility in certain regions is negatively correlated with intelligibility—regions of negative importance. While the positive correlations predominate, the negative correlations are interesting because they represent regions of the signal that are misleading to the listener. Examples of this can be found for the words /ɑdɑ/ and /ɑtʃɑ/ as the blue regions in Figs. 3(b) and 3(c). For /ɑtʃɑ/, the large vertical positive correlation is followed by a vertical negatively correlated region. This region of negative correlation might represent situations in which the listener notices a dip in high-frequency energy in the noise (a glimpse) revealing a portion of the target word in which no energy is present. This lack of speech energy in that region might be misinterpreted as a lack of high-frequency energy in the entire word, encouraging the listener to select a different, incorrect word. Thus it is possible that this negative correlation is not a function of the speech energy in that region, but instead a function of the location of that region relative to other speech energy.

C. Predictive analysis

The ability of the machine learning classifier to accurately predict whether mixtures will be correctly or incorrectly identified by human listeners is reflected in Tables IV–VII. As stated earlier, Tables IV versus V suggest a counterintuitive conclusion: The ability to accurately classify mixtures as identifiable or not is better when the classifier is trained and tested on different utterances relative to when it is trained and tested on the very same utterances. However, Table VI clarifies this result by providing classification accuracy when the number of training examples is the same in both cases. There, it can be seen that classification accuracy (the ability of the machine learning model to accurately predict whether human listeners will be able to correctly identify given mixtures) is (i) better than chance in almost every case. (ii) When the same talker is employed for training and testing of the SVM, accuracy is similar when trained and tested on the same utterances as when it is trained and tested on different utterances. (iii) However, when different talkers are employed for training and testing, accuracy is slightly superior when trained and tested on the same utterances.

The top half of Table VII shows that the classifiers are able to predict the intelligibility of mixtures significantly better than chance for all words in all conditions: All combinations of same and different talker and same and different word. In the same-word, different-talker condition, the classifiers achieved an accuracy of 59%–74%, all of which are significantly better than chance. This is due to the fact that all talkers share a common phonetic structure when producing a given word, and that time-alignment across different utterances was performed. In the two different-word conditions, the classifiers achieved an accuracy of 57%–64%, all of which are again significantly better than chance. This high classification accuracy across different training and test words is likely due to the fact that all of the utterances share the same general structure (VCV), i.e., the important regions for distinguishing between these words are all approximately aligned. However, rows five and six of the table show that the intelligibilities of mixtures involving the words /ɑdɑ/, /ɑtɑ/, and /ɑvɑ/ are predicted significantly better than the pooled cross-word results. This means that the classifiers for those words capture a significant amount of utterance-specific information beyond the general structure shared by all of the words. The bottom two rows show that /ɑfɑ/ is significantly better able to generalize across utterances from the same talker than from different talkers, indicating that only those classifiers are capturing a significant amount of talker-specific information.

V. SUMMARY AND CONCLUSION

This paper introduces an intelligibility prediction framework that is able to identify, in a data-driven manner, the importance of individual time-frequency points to the intelligibility of individual utterances. Potentially interesting observations can be made from these results that may generalize beyond the specific utterances analyzed in the current experiments. First, the absence of energy may play a larger role in understanding speech in noise than was previously assumed. In particular, a lack of energy in a key T-F region can potentially exclude the possibility of certain alternative interpretations of an utterance. The current TFIFs show that listeners focus not just on where energy does occur in choosing between possible interpretations, but also where it might occur in other interpretations. For example, although /ɑtɑ/ is unvoiced, listeners still attend to the region where voicing would occur in /ɑdɑ/ in order to correctly discriminate between the two words. While the importance of silence and targeted listening is clear in the current closed-task set, it may also transfer to open-set tasks, as the semantic context provided by the words prior and subsequent to a given target sound will serve to constrain considerably its possible identities. Similarly, the observed propensity of listeners to focus on consonant transitions into following vowels more than transitions in from preceding vowels may also very well transfer to open-set recognition.

The current study also introduces a novel intelligibility-prediction framework and shows that it is able to generalize not only to novel noise instances, but also to novel utterances of the same word from the same talker and from different talkers, and to novel consonants in /aCa/ nonsense words. These abilities serve as the necessary first steps toward creating a classifier-based intelligibility predictor. Such a model could generalize from a finite amount of training data collected from listeners, to predict intelligibility of future unseen mixtures. Whereas the current study shows that this generalization is possible across words that share a particular form, future studies are necessary to determine the extent to which it is possible for such models to generalize to the same forms in different phonetic contexts and to entirely different forms.

Going forward, the current technique permits the investigation of the optimality of listeners' strategies in various contexts. Clearly, listeners can focus only on essential disambiguating cues in more restricted closed-set tasks. Such optimality is suggested by the absence of high-frequency importance regions for /ɑfɑ/ and /ɑvɑ/ in Fig. 7. But less is known about the extent to which listeners perform this same type of optimal listening as the set of alternatives becomes larger. It is reasonable to assume that the number of time-frequency regions that must be monitored will increase substantially during open-set recognition. But it is also true that surrounding context serves to considerably limit the possible correct responses. This context is widely assumed to guide a "top-down" process that aids recognition of the bottom-up acoustic information that is gathered. What is not well understood is the extent to which listeners take advantage of this surrounding context to perform targeted (optimal) listening by modifying the bottom-up cues that are collected.

Such strategies might also help explain the ability of listeners to quickly adapt to adverse conditions such as filtering (Haggard, 1974; Darwin *et al.*, 1989) and reverberation (Watkins, 2005). Under these conditions, certain cues are systematically degraded while others are preserved, and one would expect to see a corresponding systematic shift away from degraded and toward preserved cues. The current technique permits this investigation.

The technique employed here and the resulting ability to characterize listener strategies could help to characterize deficits in noise robustness observed in various populations of listeners, such as those with dyslexia (Ziegler *et al.*, 2009), Auditory Processing Disorder (Lagacé *et al.*, 2010), and children with histories of otitis media (Zumach *et al.*, 2009). It could help characterize the large individual differences in noise robustness between hearing-impaired listeners observed even after accounting for differences in audibility (Akeroyd, 2008), perhaps in combination with detailed models of impaired hearing (e.g., Zilany *et al.*, 2009). It could be used to compare differences found between late versus early second-language learners (Mayo *et al.*, 1997), including those of specific languages over the course of their learning (e.g., Akahane-Yamada and Tohkura, 1990). It could help characterize the effects of cognitive load on listening strategy (Zekveld *et al.*, 2011). And it could help identify how musicians achieve greater noise robustness than non-musicians (Parbery-Clark *et al.*, 2009).

ACKNOWLEDGMENTS

This work was funded in part by NIH Grant No. R01 DC08594 to E.W.H.

[1] Very recently, Venezia *et al.* (2016) also extended the bubbles concept from vision research to speech recognition. The goal was not to identify regions of the speech spectrogram important for recognition of specific phonemes in background noise, but instead to identify general spectro-temporal characteristics important for intelligibility (in quiet). Rather than introducing a bubble-noise mask to the spectrogram, the authors attenuated bubble-shaped regions of the speech modulation power spectrum, which plots temporal modulation rates against spectral modulation rates. Venezia *et al.* found that modulations that contributed most to intelligibility were confined primarily to low temporal ($<$10 Hz) and low spectral ($<$2 cyc/kHz) rates, which as the authors note, supports prior work employing different techniques.

Akahane-Yamada, R., and Tohkura, Y. (**1990**). "Perception and production of syllable-initial english /r/ and /l/ by native speakers of Japanese," in *The First International Conference on Spoken Language Processing, ICSLP 1990*, Kobe, Japan (November 18–22).

Akeroyd, M. A. (**2008**). "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults," Int. J. Audiol. **47**, S53–S71.

Alcántara, J. I., Moore, B. C. J., Kühnel, V., and Launer, S. (**2003**). "Evaluación del sistema de reducción de ruido en un auxiliar auditivo digital comercial" ("Evaluation of the noise reduction system in a commercial digital hearing aid"), Int. J. Audiol. **42**, 34–42.

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

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).

ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Apoux, F., and Bacon, S. (**2004**). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," J. Acoust. Soc. Am. **116**, 1671–1680.

Apoux, F., and Healy, E. (**2012**). "Use of a compound approach to derive auditory-filter-wide frequency-importance functions for vowels and consonants," J. Acoust. Soc. Am. **132**, 1078–1087.

Apoux, F., and Healy, E. W. (**2009**). "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," Hear. Res. **255**, 99–108.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Calandruccio, L., and Doherty, K. (**2007**). "Spectral weighting strategies for sentences measured by a correlational method," J. Acoust. Soc. Am. **121**, 3827–3836.

Cohen, J. (**1960**). "A coefficient of agreement of nominal scales," Edu. Psychol. Meas. **20**, 37–46.

Cooke, M. (**2009**). "Discovering consistent word confusions in noise," in *Proceedings of Interspeech*, pp. 1887–1890.

Cooke, M. P. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Cox, D. D., and Savoy, R. L. (**2003**). "Functional magnetic resonance imaging (fMRI) 'brain reading': Detecting and classifying distributed patterns of fMRI activity in human visual cortex," Neuroimage **19**, 261–270.

Darwin, C., Denis McKeown, J., and Kirby, D. (**1989**). "Perceptual compensation for transmission channel and speaker effects on vowel quality," Speech Commun. **8**, 221–234.

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

Doherty, K. A., and Turner, C. W. (**1996**). "Use of the correlational method to estimate a listener's weighting function of speech," J. Acoust. Soc. Am. **100**, 3769–3773.

Ellis, D. (**2003**). "Dynamic time warp (DTW) in MATLAB," http://www.ee.columbia. edu/~dpwe/resources/matlab/dtw/ (Last viewed October 2, 2016).

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Gosselin, F., and Schyns, P. G. (**2001**). "Bubbles: A technique to reveal the use of information in recognition tasks," Vision Res. **41**, 2261–2271.

Haggard, M. (**1974**). "Selectivity for distortions and words in speech perception," Brit. J. Psychol. **65**, 69–83.

Healy, E. W., Yoho, S. E., and Apoux, F. (**2013**). "Band-importance for sentences and words re-examined," J. Acoust. Soc. Am. **133**, 463–473.

Lagacé, J., Jutras, B., and Gagné, J.-P. (**2010**). "Auditory processing disorder and speech perception problems in noise: Finding the underlying origin," Am. J. Audiol. **19**, 17–25.

Landis, J. R., and Koch, G. G. (**1977**). "The measurement of observer agreement for categorical data," Biometrics **33**, 159–174.

Li, F., Menon, A., and Allen, J. B. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**, 2599–2610.

Li, N., and Loizou, P. C. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.

Ma, J., Hu, Y., and Loizou, P. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.

Mayo, L. H., Florentine, M., and Buus, S. R. (**1997**). "Age of second-language acquisition and perception of speech in noise," J. Speech Lang. Hear. Res. **40**, 686–693.

Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009**). "Musician enhancement for speech-in-noise," Ear Hear. **30**, 653–661.

Rehan Akbani, Kwek, S., and Japkowicz, N. (**2004**). "Applying support vector machines to imbalanced datasets," in *European Conference on Machine Learning*, edited by J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Vol. 3201 of Lecture Notes in Computer Science (Springer, Berlin, Heidelberg), pp. 39–50.

Scharenborg, O. (**2007**). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun. **49**, 336–347.

Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.

Turner, C. W., Kwon, B. J., Tanaka, C., Knapp, J., Hubbartt, J. L., and Doherty, K. A. (**1998**). "Frequency-weighting functions for broadband speech as estimated by a correlational method," J. Acoust. Soc. Am. **104**, 1580–1585.

Varnet, L., Knoblauch, K., Meunier, F., and Hoen, M. (**2013**). "Using auditory classification images for the identification of fine acoustic cues used in speech perception," Front. Hum. Neurosci. **7**, 865.

Venezia, J. H., Hickok, G., and Richards, V. M. (**2016**). "Auditory 'bubbles': Efficient classification of the spectrotempoal modulations essential for speech intelligibility," J. Acoust. Soc. Am. **140**, 1072–1088.

Watkins, A. J. (**2005**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118**, 249–262.

Yu, C., Wójcicki, K. K., Loizou, P. C., Hansen, J. H. L., and Johnson, M. T. (**2014**). "Evaluation of the importance of time-frequency contributions to speech intelligibility in noise," J. Acoust. Soc. Am. **135**, 3007–3016.

Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2011**). "Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response," Ear Hear. **32**, 498–510.

Ziegler, J. C., Pech-Georgel, C., George, F., and Lorenzi, C. (**2009**). "Speech-perception-in-noise deficits in dyslexia," Dev. Sci. **12**, 732–745.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. (**2009**). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," J. Acoust. Soc. Am. **126**, 2390–2412.

Zumach, A., Gerrits, E., Chenault, M. N., and Anteunis, L. J. C. (**2009**). "Otitis media and speech-in-noise recognition in school-aged children," Audiol. Neuro-otol. **14**, 121–129.