LEARNING AN INTELLIGIBILITY MAP OF INDIVIDUAL UTTERANCES

Michael I. Mandel

Department of Computer Science and Engineering
The Ohio State University
mandelm@cse.osu.edu

ABSTRACT

Predicting the intelligibility of noisy recordings is difficult and most current algorithms only aim to be correct on average across many recordings. This paper describes a listening test paradigm and associated analysis technique that can predict the intelligibility of a specific recording of a word in the presence of a specific noise instance. The analysis learns a map of the importance of each point in the recording's spectrogram to the overall intelligibility of the word when glimpsed through "bubbles" in many noise instances. By treating this as a classification problem, a linear classifier can be used to predict intelligibility and can be examined to determine the importance of spectral regions. This approach was tested on recordings of vowels and consonants. The important regions identified by the model in these tests agreed with those identified by a standard, non-predictive statistical test of independence and with the acoustic phonetics literature.

Index Terms— Intelligibility, Noise, Glimpse, Objective, Subjective

1. INTRODUCTION

This paper describes a novel listening test that measures the intelligibility of mixtures of the same exact recording of a target word or syllable with many different instances of noise. The noise, which we call "bubbles" is designed to produce "glimpses" [1] of the speech in various regions of the spectrogram, but obliterate the speech elsewhere. This is an idealized model of the glimpsing concept, which states that listeners can reconstruct speech from noisy observations by focusing on the content of and relationship between localized regions of uncorrupted speech in the observation. We hypothesize that some glimpses are more important than others and investigate the relative importance of glimpses of particular utterances. Glimpses that are important to an utterance's intelligibility will frequently occur in intelligible mixtures and be absent from unintelligible mixtures. The locations of these important glimpses form an *intelligibility map* of the utterance.

The problem of predicting whether a mixture of speech and noise is intelligible based on features extracted from the exact speech and noise signals involved is a novel classification problem that forms the basis of this technique. We solve this classification problem using linear support vector machines (SVMs), allowing us to interpret the learned classifier as the importance of each time-frequency point, i.e., the intelligibility map. By treating "microscopic" intelligibility prediction as a classification problem in this way, we can measure the ability of the model to generalize to new mixtures using classic cross-validation procedures. By performing this experiment directly on the signals in the time-frequency domain, we can make much more specific predictions than models that are mediated by averages over signals, utterances, or time. We also perform an analysis of these same data using descriptive statistics that are more flexible in producing intelligibility maps, but cannot be used to make predictions about unseen mixtures.

2. BACKGROUND

The work in [4] investigated the importance of individual frequency bands to the intelligibility of words in the context of randomly selected subsets of other frequency bands. It shows that while this importance is quite stable across listeners, it varies a great deal across recordings. These data suggest that when little linguistic context is available, the important frequencies correspond to the frequencies of the first three formants found in those utterances. This paper expands this technique to include importance in timing information in addition to frequency, providing a much more localized and context-dependent prediction of intelligibility. The proposed procedure is the acoustic analog of the visual "bubbles" method described by [2] for identifying the regions of facial images that are important to human viewers in the visual tasks of identifying gender, expressivity, and identity. Our method is also inspired by recent advances in the field of fMRI analysis, which in the past decade have started to use classifiers to distinguish between brain states [3].

There are two systems that are similar to parts of the proposed work. The model of [5] included a number of ad-hoc time-frequency importance functions that improve the predictions of a number of existing intelligibility models in time-varying noise. In contrast, by framing the problem of intelligibility prediction as classification, the current work can learn these importance functions from data.

Perhaps the most similar to the proposed work is [6] and related papers, e.g., [7, 8], which describe the three-dimensional deep search (3DDS) technique to identify important spectral regions of individual utterances. They perform three different operations (truncation, filtering, and the addition of white noise) to varying degrees on individual utterances. The points at which these operations transform the original syllable into another syllable allow the authors to identify small regions of the spectrogram that are crucial to the identification of the particular phones. This is in some ways a new take on the classic study of acoustic phonetics, e.g., [9].

Our listening test technique differs from [6] in that by posing intelligibility prediction as classification, our models will be able to generalize to new conditions. We can use techniques like cross-validation to measure the ability of the classifiers to generalize to new noises, new listeners, and new speakers. Our ultimate goal is, after using tools from ASR to align utterances, both to measure the classifiers' ability to generalize to new utterances and to develop a microscopic intelligibility predictor that is applicable to all speech.

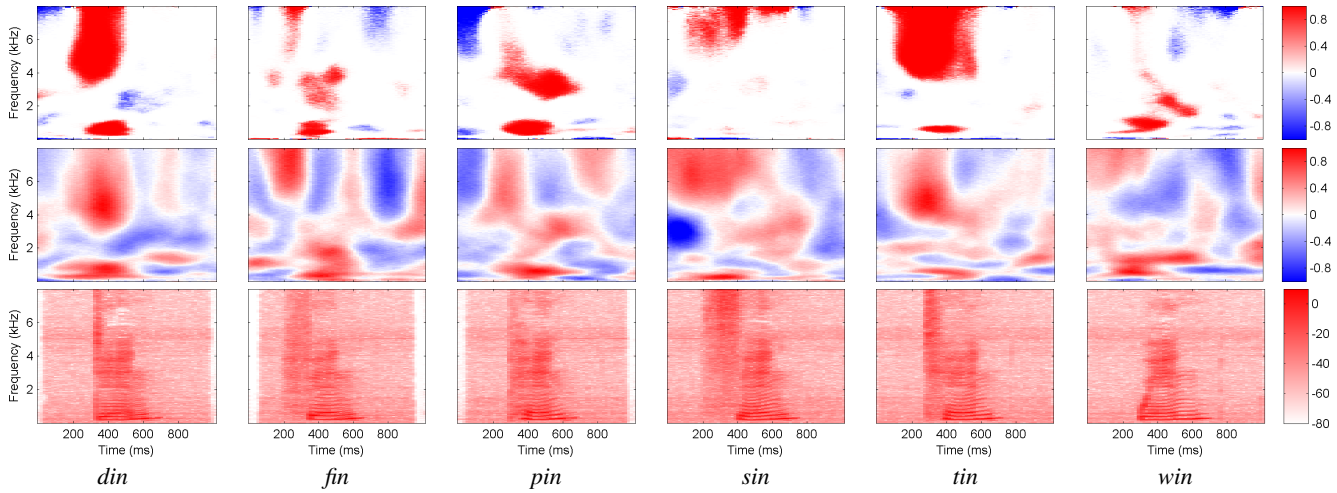

Figure 1: Statistical (top) and support vector machine-based (middle) intelligibility maps for mechanical Turk listeners on *Cin* utterances along with the spectrograms of the clean utterances (bottom).

3. LISTENING TEST

The basic setup of the listening test is a six-alternative forced choice procedure. Listeners were presented with a mixture of a single word with noise and asked to select the correct word from six choices. The speech was selected from the Hillenbrand vowel corpus [10] and the audio portion of the "CMU Audio-Visual Profile and Frontal View" (CMU AVPFV) database [11], which is an instance of the Modified Rhyme Test (MRT). The vowels were presented in the form of *hVd* syllables and consisted of *had*, *head*, *heed*, *hid*, *hood*, and *who'd* from male speaker M06. The consonants from the MRT we used were presented in the form of *Cin* syllables and consisted of *din*, *fin*, *pin*, *sin*, *tin*, and *win*. We used female speaker "Helen" and repetitions 2, 3, and 5 of the 10 to measure the consistency of our analysis over different instances of the same word from the same speaker. The MRT utterances had a rather high noise floor, as can be seen in the spectrograms in Figures 1 and 3, but this did not affect their intelligibility. All speech was sampled at 16 kHz.

The noise was designed to provide glimpses of the speech only in specific time-frequency bubbles. It started with speech-shaped noise loud enough to completely overwhelm the speech. We found that an SNR of -28 dB was sufficient to make the MRT speech completely inaudible and -33 dB for the Hillenbrand vowels. We then attenuated this noise in "bubbles" that were jointly parabolic in time and ERB-scale frequency [12] with a maximum suppression of 80 dB. The centers of the bubbles were selected uniformly at random in time and in ERB-scale frequency, except for a 2-ERB buffer at both the bottom and top of the frequency scale. Mathematically, the mask applied to the speech-shaped noise, $M(f,t)$, is

$$B(f,t) = \sum_i \exp\left\{-\frac{(t-t_i)^2}{\sigma_t^2} - \frac{(E(f)-E(f_i))^2}{\sigma_f^2}\right\} \quad (1)$$

$$M(f,t) = \min\left(1, \frac{10^{-80/20}}{B(f,t)}\right) \quad (2)$$

where $E(f) = 21.4\log_{10}(0.00437f + 1)$ converts frequencies in Hz to ERB, and $\{(f_i,t_i)\}_{i=1}^{I}$ are the centers of the I bubbles. We set σ_t and σ_f such that the bubbles are 350 ms wide at their widest and 7 ERB high at their highest, which were the smallest values that would avoid introducing audible artifacts. The number of bubbles per second was set such that listeners could correctly identify approximately 50% of the mixtures. We found that 12 bubbles per second achieved this level for the MRT words and 10 bubbles per second for the Hillenbrand vowels.

We used two groups of participants. The first group consisted of a single expert listener who was familiar with the clean speech material. The second was a collection of 74 workers on Amazon's mechanical Turk website[1]. With IRB approval, these listeners provided their informed consent to participate in the experiment and stated that they were at least 18 years of age with normal hearing.

We created 199 mixtures of each word with randomly generated bubble noise. We found that this number of mixtures was necessary for reliable analyses. The vowel and MRT tasks were performed separately, so listeners heard mixtures of only *hVd* words or *Cin* words and selected from the corresponding list. The noisy mixtures were interspersed with clean versions of the words to ensure that the listeners were paying attention and to familiarize them with the words. These noise-free utterances constituted between 10 and 20% of the utterances presented to listeners, but were not included in the main data analysis. Listeners were informed that if they incorrectly answered more than 33% of these clean utterances their work would be rejected, but this situation never arose. The mechanical Turk listeners also listened to noise free versions of these utterances 15 times each to ensure that the original speech was intelligible to them.

Workers on mechanical Turk were paid \$0.08 to identify 5 mixtures in a single "Human Intelligence Task" (HIT). On average, each of the 74 participants completed 21 HITs, with the median being 6 HITs. Five of the participants completed more than 130 HITs each. The median time to complete one HIT was 27 seconds, leading to a median hourly rate of \$10.67. This rate is comparable to subjects recruited in person, but we were able to collect data much more quickly, for example, labeling 3000 mixtures in 2 hours. The total cost of running these experiments, including pilot tests was \$250.

We experimented with different numbers of listeners per mixture. The expert listener heard 12 repetitions of nine mixtures of each *Cin*. That listener's results were self-consistent in that the same mixture

[1] http://mturk.com

generated the same response nearly all of the time. The mechanical Turk listeners heard these same mixtures 21 times each. Their results were less consistent, with the percentage correct for each mix being nearly uniformly distributed between 0 and 100%. It is not clear whether this is due to differences in skill, familiarity with the speech material, or another cause. The rest of the mixtures were heard either by 1 or 2 listeners, as shown in the $\times$ column of Table 1.

4. INTELLIGIBILITY MAPS

We analyzed the intelligibility of these various mixtures to determine whether there are certain regions of the spectrogram of each word that must be "glimpsed" for a particular mixture to be intelligible. A mixture was considered intelligible if at least 70% of listeners correctly identified it, unintelligible if at most 30% of listeners did, and excluded from the analysis if it fell between these two values.

4.1. Descriptive: Statistical testing

The first intelligibility map estimator we propose is a descriptive statistical test using Pearson's chi-squared test of independence for each time-frequency point. This test determined whether or not the number of times that a particular point was glimpsed was independent of the number of times the mixture was intelligible. If the test revealed that with high confidence these two variables were not independent at a particular point, then that point probably contributed to the intelligibility of the word, either positively or negatively.

Specifically, we computed the contingency table, $C_{ij}(f,t)$, for each clean word, which contained the count of the number of mixtures in which point (f,t) was included or not included in a bubble ($i = 1, 0$, respectively) and in which the mixture was intelligible or unintelligible ($j = 1, 0$, respectively). The test statistic is then

$$X^2(f,t) = \sum_{i,j=0}^{1} \frac{(C_{ij}(f,t) - E_{ij}(f,t))^2}{E_{ij}(f,t)} \quad (3)$$

where if N is the total number of mixtures,

$$E_{ij}(f,t) = \frac{1}{N}\Big(\sum_{i=0}^{1} C_{ij}(f,t)\Big)\Big(\sum_{j=0}^{1} C_{ij}(f,t)\Big). \quad (4)$$

We then computed $P(f,t)$, the likelihood of $X^2(f,t)$ under a χ^2 distribution with 1 degree of freedom. If $P(f,t)$ is less than $\alpha = 0.05$ then i and j are probably not independent. If $C_{11}(f,t) > E_{11}(f,t)$ then we assume that that time-frequency point contributes to intelligibility. Otherwise we assume it contributes against intelligibility. We visualize this as an intelligibility map

$$M_\chi(f,t) = \text{sign}\left(C_{11}(f,t) - E_{11}(f,t)\right) \exp\left(-\tfrac{1}{\alpha}P(f,t)\right) \quad (5)$$

an example of which is shown in the top row of Figure 1.

4.2. Predictive: Support vector machine

The second method for computing an intelligibility map used a linear classifier, in this case we used a support vector machine (SVM). This method is predictive because, in contrast to the descriptive method, it allows the quality of the fit to be measured via the prediction accuracy of the model on unseen data.

The mixtures of each clean recording constituted a single learning problem. The features that we used are $G_m(f,t)$, the amount that the speech shaped noise has 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, the 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 $\times$ 64 frames = 32832 dimensions), we first performed principal components analysis (PCA) to reduce this dimensionality to between 5 and 120. Computing PCA on the features directly gave too much weight to the high frequency bubbles, which, for the same number of ERBs include many more short-time Fourier transform (STFT) frequency channels than the low frequency bubbles. We thus re-weighted the features before performing PCA to counteract this effect. The weight we used was the cube root of the incremental ERB frequency change between adjacent STFT frequency channels. We used nested cross-validation to select the best PCA dimensionality on the training data of each cross-validation fold. The dimensionality selected was generally between 12 and 31, with 31 being the most common by a small margin. The visualizations in Figures 1 and 3 use 31 PCA dimensions and are trained on all of the data available. If the linear classifier is

$$\hat{y}_m = b + \sum_k w_k \sum_{f,t} B_k(f,t) G_m(f,t) \quad (6)$$

where $B_k(f,t)$ is the kth PCA basis, then the corresponding intelligibility map is $M_s(f,t) = \sum_k w_k B_k(f,t)$. Because its predictions are invariant to its scale, the visualizations show $M_s(f,t)$ after scaling its maximum value to 1.

Table 1: Nested cross-validation accuracy of support vector machine classifiers on expert (Exp) and mechanical Turk (MT) listeners for *hVd* words and three different spoken instances of *Cin* words (Ver). Each row used 199 mixtures of each word and certain mixtures were heard by multiple people ($\times$).

			Word					
	Ver	$\times$	*din*	*fin*	*pin*	*sin*	*tin*	*win*
Exp	1	1	73.1	70.7	72.5	64.9	75.9	58.8
MT	1	2	76.6	68.1	81.5	52.1	75.6	71.9
MT	2	1	70.5	58.5	51.0	61.5	72.6	51.8
MT	3	1	77.2	57.3	55.0	61.9	68.9	57.5
			Word					
	Ver	$\times$	*had*	*head*	*heed*	*hid*	*hood*	*who'd*
Exp	1	1	70.0	70.4	55.6	61.3	57.2	71.5
MT	1	1	67.2	58.8	59.6	55.7	52.7	56.4

5. ANALYSIS

Table 1 shows the accuracy of the SVM in predicting the intelligibility for each of the original speech utterances. These accuracies were measured on test sets with equal numbers of positive and negative examples, balanced separately on each cross-validation fold, so the baseline accuracy is 50%. They were averaged over three repetitions of nested 10-fold cross-validation of 199 mixtures. On the consonants, the table shows that the accuracies are quite good for the expert listener and the mechanical Turk listeners on the first instances. The other two instances were only seen by a single mechanical Turk listener each (denoted by the $\times$ column) and so, while

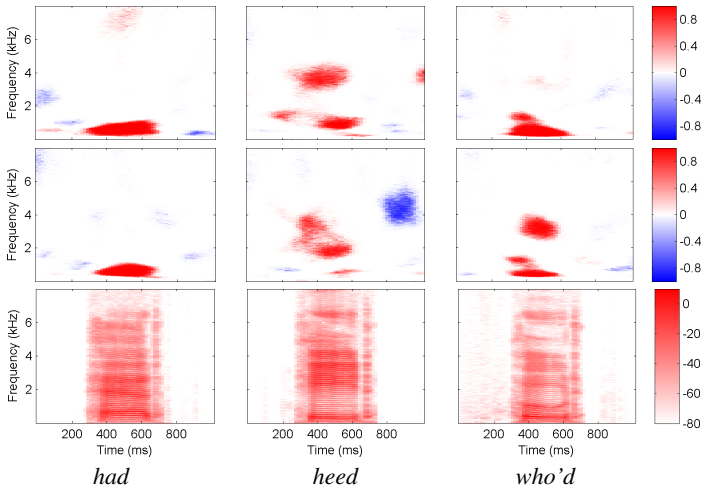

Figure 2: Statistical intelligibility maps for expert (top) and mechanical Turk (middle) listeners on *hVd* utterances, with spectrograms of the clean utterances (bottom).

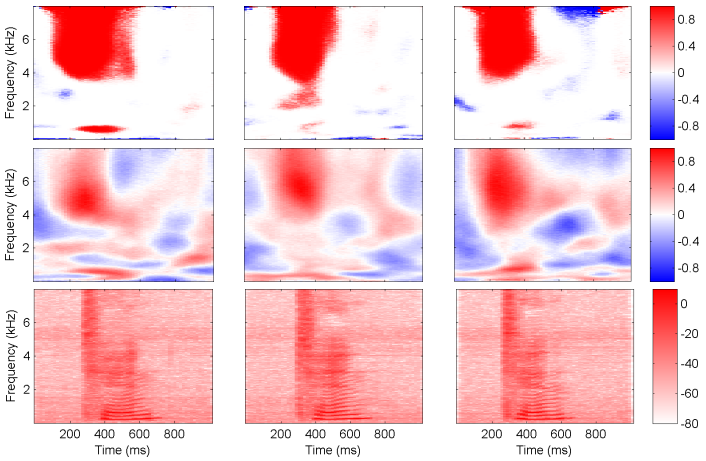

Figure 3: Intelligibility maps for mechanical Turk listeners across three instances of *tin* spoken by the same speaker: statistical (top), SVM (middle), spectrograms of the clean utterances (bottom).

still learnable for *din* and *tin*, are less learnable for the other consonants. The vowels were similarly better learned for the expert listener than for the single exposure mechanical Turk listeners.

Figure 1 shows examples of the two intelligibility maps for the first set of the *Cin* utterances based on the mechanical Turk data. Notice how the intelligibility maps tend to agree with standard theories of acoustic phonetics. Specifically, the important regions for the plosive consonants /d/ and /t/ in *din* and *tin* are the high frequencies just before the vowel for detecting the burst and the simultaneous low frequencies for detecting voicing. The plosive /p/ shows similar trends, although to a lesser degree. The fricatives /s/ and /f/ in *sin* and *fin* both place high importance on high frequencies before the vowel, with the region lower and longer temporally for *sin* than *fin*. The semi-vowel /w/ in *win* places importance on the audibility of the first formant and the dipping second formant at the beginning of the syllable. The SVM map provides more detail than the statistical map, although the statistical map permits absolute judgments of importance.

Figure 2 shows the statistical intelligibility maps for the expert and mechanical Turk listeners on the *hVd* utterances, which are very similar. The intelligibility maps of these vowels are quite different from those of the consonants. They generally only depend on the first formant being audible or the peak and higher frequency roll-off of the first formant. The word *who'd* also places some emphasis on the second formant, especially at the beginning of the syllable.

The variation of the intelligibility map for the single word *tin* across recordings is shown in Figure 3. While there is some variability between the three instances, all of them contain the main pattern of a high frequency burst and a low frequency voicing detection. As shown in Table 1, accuracies for all three instances are high.

6. CONCLUSION

This paper has described a listening test paradigm and associated analysis technique that can predict the intelligibility of a specific recording of a word in the presence of a specific noise instance. By treating this as a classification problem, a linear classifier can be used to predict intelligibility and can be examined to determine the importance of spectral regions. In testing the approach on recordings of vowels and consonants, the important regions identified by the model agreed with those identified by a descriptive statistical test and with the acoustic phonetics literature. Subsequent work will generalize this model to predict the intelligibility of unseen utterances in noise, to compare automatic speech recognition in noise to human speech recognition, and to evaluate speech processing algorithms' preservation of spectral regions that are important to intelligibility.

7. REFERENCES

[1] M. P. Cooke, "A glimpsing model of speech perception in noise," *J. Acous. Soc. Am.*, vol. 119, pp. 1562–1573, 2006.

[2] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261–2271, 2001.

[3] J. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nat. Rev. Neurosci.*, vol. 7, no. 7, pp. 523–534, 2006.

[4] E. W. Healy, S. E. Yoho, and F. Apoux, "Band importance for sentences and words reexamined," *J. Acous. Soc. Am.*, vol. 133, no. 1, pp. 463–473, 2013.

[5] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band importance functions," *J. Acous. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[6] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *J. Acous. Soc. Am.*, vol. 127, no. 4, pp. 2599–2610, 2010.

[7] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *J. Acous. Soc. Am.*, vol. 121, no. 4, pp. 2312–2326, 2007.

[8] J. Allen and F. Li, "Speech perception and cochlear signal processing," *Signal Processing Magazine, IEEE*, vol. 26, no. 4, pp. 73–77, 2009.

[9] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics. MIT Press, 2000.

[10] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *J. Acous. Soc. Am.*, vol. 97, no. 5, pp. 3099–3111, 1995.

[11] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *Proc. IEEE ICASSP*, vol. 4, Honolulu: IEEE, 2007, pp. 429–432.

[12] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.