Generalizing time-frequency importance functions across noises, talkers, and phonemes

Michael I Mandel[1], *Sarah E Yoho*[2], *and Eric W Healy*[2]

[1]Computer Science & Engineering, The Ohio State University, USA
[2]Speech & Hearing Science, The Ohio State University, USA

mandelm@cse.osu.edu, yoho.17@osu.edu, healy.66@osu.edu

Abstract

Listeners can reliably identify speech in noisy conditions, although it is generally not known what specific features of speech are used to do this. We utilize a recently introduced data-driven framework to identify these features. By analyzing listening-test results involving 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. This paper shows that a trained model resulting from this framework can generalize to new conditions, successfully predicting the intelligibility of novel mixtures. First, it can generalize to novel noise instances after being trained on mixtures involving the same speech utterance but different noises. Second, it can generalize to novel talkers after being trained on mixtures involving the same syllables produced by different talkers in different noises. Finally, it can generalize to novel phonemes, after being trained on mixtures involving different consonants produced by the same or different talkers in different noises. Aligning the clean utterances in time and then propagating this alignment to the features used in the intelligibility prediction improves this generalization performance further.

Index Terms: speech in noise, intelligibility prediction, glimpses, generalization

1. Introduction

Normal-hearing listeners are remarkably good at understanding speech in noisy environments, much better than hearing-impaired listeners [1, 2] and automatic speech recognition systems [3–6]. Understanding this robustness and reproducing it in machine listeners would likely enable vast improvements in hearing aids and conversational interfaces. One promising theory of the mechanism underlying this process is that listeners detect relatively clean "glimpses" of speech in the acoustic signal and assemble them into a hypothesized utterance [3, 7–9]. We have recently shown that some glimpses are more useful than others for correctly identifying a particular utterance in noise [10], yielding a determination of "where" in the speech signal listeners find noise robust linguistic information.

Specifically, in [10] we describe a listening test and two analyses for identifying time-frequency regions of individual utterances that are important for those utterances to be correctly identified in noisy mixtures. The human listening test measured the intelligibility of mixtures of the same exact utterance with many different instances of "bubble" noise, which approximated randomized glimpsing. A *descriptive analysis* was able to identify important time-frequency regions from these data. A *predictive analysis* was able to learn to predict the intelligibility of particular mixtures, i.e., to generalize across noise instances.

This paper measures the ability of that framework to generalize across different utterances. Specifically, we show that it is able to generalize to new instances of the same word spoken by the same talker and by different talkers. We also show that these models are able to predict better than chance the intelligibility of somewhat different words by virtue of the words' similarity to one another. In order to improve the ability of these classifiers to make these predictions, we introduce a technique to align the clean utterances using dynamic time warping (DTW) and propagate this alignment into the features used by the classifiers.

2. Background

This work is inspired by methods from several fields. In [11], we measured the band importance function (BIF) for several speech datasets, and found that these functions were very consistent across listeners, but very dependent on the speech material under analysis. These data imply that importance may be a property of the speech itself, not a global property of the frequency spectrum. With this in mind, we introduce here the time-frequency importance function (TFIF), a weighting of the importance of each time-frequency point in an utterance to its intelligibility. A TFIF is essentially a time-varying and utterance-specific BIF.

Ma *et al.* [12] showed that ad-hoc time-frequency weighting functions can improve the performance of objective predictors of speech intelligibility. We believe that our data-driven TFIFs should improve these predictions even more. Li *et al.* [13] introduced 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. While this technique can only be applied to initial and final phonemes of utterances, our technique can be applied to phonemes in any position in a word, even in the context of running sentences.

The "audio bubbles" listening test methodology is based on the visual bubbles test [14], which uses a visual discrimination task to identify the regions of images important for viewers to identify expressivity, gender, and identity. Recently, [15] proposed an application of the Classification Image, a related technique, to the auditory task of discriminating /b/ from /d/. Their use of white Gaussian noise as the corruption required more than 3000 mixtures of each token to generate a reliable analysis. Our audio bubbles require an order of magnitude fewer mixtures to perform a similar task by exploiting the large variation in importance of different "glimpses" of the clean speech.

3. Listening test

The speech material was selected from the corpus described in [16]. Two different experiments were performed using two different sets of utterances. Both used several pronunciations of six vowel-consonant-vowel (VCV) nonsense words. The nonsense words were of the form *aCa*: *acha*, *ada*, *afa*, *aja*, *ata*, *ava*. Experiment 1 used three utterances of each word from a single female talker (number W3). We selected the slowest and fastest versions of each utterance from approximately 10 versions along with one of an intermediate duration. Experiment 2 used one version of each utterance from three different talkers, numbers W2, W4, and W5. These talkers were selected because their recordings were of the highest quality and they showed a large variation in speaking style. We selected talkers of the same gender so that they had similar pitches and formant positions. We selected female talkers because they had fewer pronunciation mistakes than the male talkers. The utterances were all 2.2 seconds long including surrounding silence.

Each utterance was mixed with 200 instances of "bubble" noise. This noise was designed to provide glimpses of the speech only in specific time-frequency bubbles. To construct this noise, we began with speech shaped noise with an SNR of -24 dB, sufficient to make the speech completely unintelligible. The noise was then attenuated in "bubbles" that were jointly parabolic in time and ERB-scale frequency [17]. The center points of the bubbles were selected uniformly at random in time and in ERB-scale frequency, except that they were excluded from a 2-ERB buffer at the bottom and top of the frequency scale to avoid edge effects. Mathematically, one instance of bubble noise is given by

$$B(f,t) = \sum_{i=1}^{I} \exp\left\{ -\frac{(t-t_i)^2}{\sigma_t^2} - \frac{(E(f)-E(f_i))^2}{\sigma_f^2} \right\}$$
$$N(f,t) = \gamma S(f) \min\left(1, \frac{10^{-\eta/20}}{B(f,t)}\right) \tag{1}$$

where $E(f) = 21.4 \log_{10}(0.00437f + 1)$ converts frequencies in Hz to ERB, $\{(f_i, t_i)\}_{i=1}^{I}$ are the randomly selected centers of the I bubbles, γ controls the overall gain, and $S(f)$ is the average spectrum of all of the clean utterances used in the experiments. We used a maximum suppression of $\eta = 80$ dB and set σ_t and σ_f such that the bubbles were 350 ms wide at their widest and 7 ERB high at their highest, the smallest values that would avoid introducing audible artifacts. The number of bubbles was set such that listeners could correctly identify approximately 50% of the mixtures, which we found to be 15 bubbles per second.

Subjects were 8 volunteers having normal hearing as defined by audiometric thresholds on day of test ≤ 20 dB HL at octave frequencies from 250 to 8000 Hz [18,19]. They were aged 18–22 and participated for extra course credit. They 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 and Sennheiser HD280 PRO headphones. Sound presentation levels were calibrated via Larson Davis sound level meters and couplers so that mixtures were presented at 75 dBA.

Each listener was assigned to either Experiment 1 or 2 and was assigned 900 mixtures, 50 involving each of the 18 VCVs from the relevant set. Each mixture was heard by a single listener. 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 a closed set of six using a textual MATLAB interface. There were two 5-minute training periods where the listener was given feedback on their responses. The first used the clean utterances and the second used noisy utterances. No feedback was given during the main listening test, which took approximately one hour per subject.

Talker	*acha*	*ada*	*afa*	*aja*	*ata*	*ava*
W3 v1	72.0	60.5	60.0	70.5	75.0	70.5
W3 v2	72.0	64.5	64.5	69.5	72.0	76.5
W3 v3	74.0	67.5	57.0	64.5	50.0	69.5
W2	32.0	52.5	32.0	49.5	58.5	75.0
W4	75.0	75.5	64.0	54.5	77.0	56.0
W5	52.5	62.5	51.0	47.0	73.5	67.5

Table 1: Percent of mixtures correctly identified by listeners

The percentage of mixtures that were correctly identified for each clean utterance is shown in Table 1. Note that many of these percentages are above our target value of 50%. As discussed in Section 4.2, the effect of this is to reduce the number of mixtures per utterance that can be utilized in some of our analyses.

4. Analyses

We utilize the descriptive and predictive analyses of [10]. Both analyses are performed on features computed from each mixture, $\tilde{N}_j(f,t) = 20 \log_{10} N_j(f,t)$, the noise level at each point in the spectrogram, and y_j, the proportion of times the mixture was identified as containing the correct word. Note that these features do not represent the speech, which is only represented implicitly through groupings of related mixtures. We also introduce here a procedure to align two clean utterances and propagate that alignment to the corresponding features.

4.1. Descriptive analysis: Statistical testing

The descriptive analysis performs Pearson's χ-squared test of independence at each spectrogram point between $\tilde{N}_j(f,t)$ and y_j. Points that are *not* independent of the mixture's intelligibility are identified as important to it. The time-frequency importance function (TFIF) is a matrix with one point corresponding to each time-frequency point in an utterance's spectrogram, and a value at that point reflecting its importance,

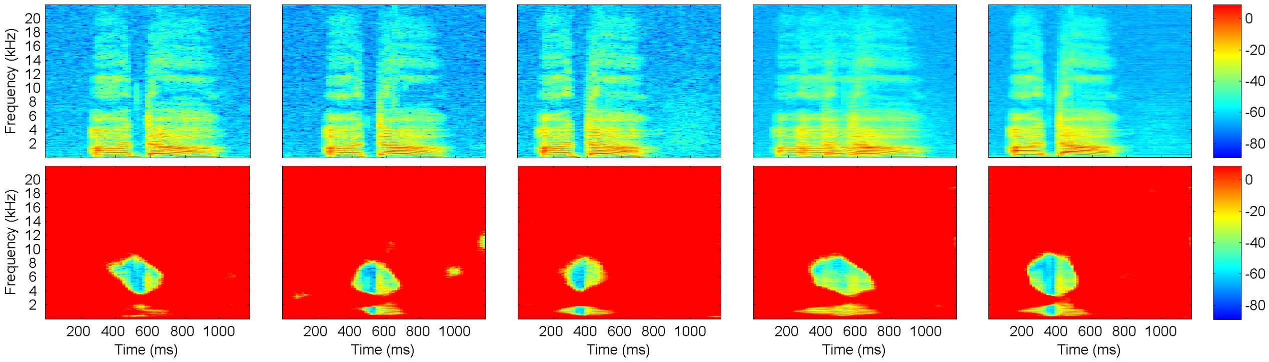
$$M_\chi(f,t) = \pm \exp\left(-\tfrac{1}{\alpha} P(f,t)\right) \tag{2}$$

where $P(f,t)$ is the p-value of the χ-squared test, α is a significance level of 0.05, and the sign is the same as that of the correlation. This analysis is similar to the correlational method of identifying band importance functions from listening tests [20].

4.2. Predictive analysis: Support vector machine

The predictive analysis is performed by a linear support vector machine (SVM), selected because of its interpretable representation and good generalization performance. We train the model to predict y_j from $\{\tilde{N}_j(f,t)\}_{\forall f,t}$. The dimensionality of these features is very high (approximately 100,000), so we project them into a 70-dimensional subspace using principal components analysis before feeding them into the SVM.

Because SVMs are sensitive to imbalances in the number of positive and negative training examples [21], we balanced the classes by discarding examples from whichever class is more prevalent. Thus, the effective number of training examples was twice the number in the smaller class. This meant that conditions that differed substantially from 50% correct reduced the effective number of training examples because of the resulting imbalance between intelligible and unintelligible mixtures. We balanced the number of test utterances in the same way to enforce a baseline accuracy of 50%.



(a) Talker W3 v1 (b) Talker W3 v2 (c) Talker W3 v3 (d) Unwarped (e) Warped

Figure 1: Spectrograms (top row) and time-frequency importance functions masking the spectrograms (bottom row) for three versions of the word *ada* spoken by the *same* talker (a)–(c), and the pooled spectrogram and TFIF unwarped (d) and warped (e) to match W3 v3.

Talker	*acha*	*ada*	*afa*	*aja*	*ata*	*ava*
W3 v1	54.5	**58.9**	53.8	**65.3**	49.0	54.2
W3 v2	58.0	**65.5**	58.5	**68.0**	**62.5**	59.6
W3 v3	**64.4**	58.5	**64.0**	**66.2**	**66.5**	56.6
W2	57.8	**73.2**	52.3	49.5	**62.7**	55.0
W4	60.0	54.1	57.6	**61.0**	**71.7**	**58.5**
W5	**59.5**	56.7	52.6	**61.2**	**64.2**	53.8
Avg	**59.0**	**61.1**	**56.4**	**61.9**	**62.8**	**56.3**

Table 2: 5-fold cross-validation accuracy within individual utterances. Bold entries are significantly above chance performance (50%) at a 0.05 level according to a one-sided binomial test.

4.3. Alignment: Dynamic time warping

Because the features used in the various analyses only represent the clean speech implicitly, some processing was necessary in order to permit generalization between different utterances. Various normalizations, including pitch, vocal tract length, and timing, could be performed between clean utterances and then propagated to the features of the relevant mixtures. Timing normalization is the most straightforward to propagate to the features, so that was the only normalization that we used here. Specifically, a time warp for a clean utterance was computed using [22] to minimize the sum of squared errors between its mel frequency cepstral coefficients and those of a target clean utterance, with no penalty for insertions or deletions. This warp was then applied to the features of the first utterance's mixtures before performing further analyses.

5. Results

Figure 1 shows the spectrograms and TFIFs of three different versions of the word *ada* spoken by a single talker. It also shows a pooled spectrogram, computed by averaging the magnitude spectrograms of all of the individual utterances, and the pooled TFIF both before and after warping the utterances to align with W3 v3. It shows that the original utterances are severely misaligned, causing no problem for listeners, but disrupting the naïve analysis. Aligning them leads to more consistent results.

Figure 2 shows the spectrograms and TFIFs of the word *ada* spoken by four different talkers (including one utterance from the talker in Figure 1). It shows that such utterances exhibit greater variation than those from the same talker. Aligning them again helps make them more comparable to one another and the results more consistent.

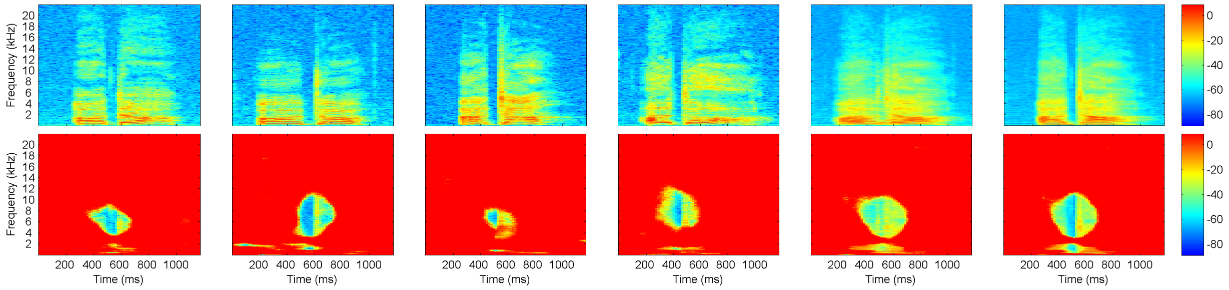
Talker	Warp	*acha*	*ada*	*afa*	*aja*	*ata*	*ava*
Same	+	65.2	74.1	67.0	62.7	74.4	63.5
Same	−	63.6	63.4	57.2	63.9	74.0	64.5
Diff	+	62.3	69.3	59.9	58.7	73.9	61.6
Diff	−	55.7	66.8	61.6	58.9	63.1	59.9

Table 3: Cross-utterance classification accuracy on mixtures involving a novel utterance. 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.

5.1. Predictive analysis

Table 2 serves as a baseline for subsequent SVM accuracy results. It shows the accuracy of classifiers trained and tested on mixtures that involve the same clean speech utterance using 5-fold cross-validation. Thus 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. It shows the accuracy that classifiers can achieve when required to generalize across noise instances and not across speech utterances. It is limited, however, by the amount of training data available for each classifier, at most 160 mixtures per utterance. Table 4 (discussed below) will show that increasing the amount of training data, even from different utterances, can improve this accuracy significantly. We would expect, therefore, that these cross-validation numbers would continue to improve given additional mixtures for each utterance.

Table 3 shows the cross-utterance classification accuracy of the SVM classifiers. It shows generalization results for three different utterances from the same talker and for four different utterances from four different talkers. In each case, a model was tested on a single utterance after being trained on the other utterances. This procedure was repeated, rotating through each of the utterances for testing, with the accuracies averaged together. The table shows these results for both the unmodified features and features aligned to W3 v3 for the single-talker condition and W4 for the multiple talker condition using the technique described in Section 4.3. 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 aligned to a reference. Subsequent results will therefore only be reported with the use of warping.



(a) Talker W3 v1 (b) Talker W2 (c) Talker W4 (d) Talker W5 (e) Unwarped (f) Warped

Figure 2: Spectrograms (top row) and time-frequency importance functions masking the spectrograms (bottom row) for four versions of the word *ada* spoken by *different* talkers (a)–(d) , and the pooled spectrogram and TFIF unwarped (e) and warped (f) to match W4.

T	U	N_{tr}	*acha*	*ada*	*afa*	*aja*	*ata*	*ava*
S	S	104.9	**59.0**	**60.9**	**58.7**	**66.5**	**59.3**	**56.8**
S	D	105.1	**57.6**	**69.5**	**56.5**	**62.0**	**63.7**	**60.0**
S	D	262.2	**65.2**	**74.1**	**67.0**	**62.7**	**74.4**	**63.5**
D	S	114.3	**57.9**	**60.7**	**54.1**	**59.2**	**61.9**	**55.4**
D	D	114.1	50.7	**54.8**	**55.1**	49.6	**59.0**	53.9
D	D	464.7	**62.3**	**69.3**	**59.9**	**58.7**	**73.9**	**61.6**

Table 4: Accuracy of cross-utterance classification and cross-validation within utterances showing the effect of increased training data. Abbreviations: Talker (T), Utterance (U), Same (S), Different (D), and number of training points (N_{tr}). Bold entries are significantly better than chance at a 0.05 level according to a one-sided binomial test.

Talker	Word	*acha*	*ada*	*afa*	*aja*	*ata*	*ava*
Same	Same	**65.2**	**74.1**	**67.0**	**62.7**	**74.4**	**63.5**
Same	Diff	**62.2**	**61.8**	**61.9**	**59.8**	**57.5**	**56.8**
Diff	Same	**62.3**	**69.3**	**59.9**	**58.7**	**73.9**	**61.6**
Diff	Diff	**64.0**	**61.4**	**59.9**	**62.0**	**58.0**	**59.4**
Same	Δ	3.0	**12.3**	5.2	2.8	**16.9**	**6.7**
Diff	Δ	-1.8	**7.9**	0.0	-3.3	**15.8**	2.2
Δ	Same	2.9	4.9	**7.1**	4.0	0.6	1.9
Δ	Diff	-1.8	0.5	2.0	-2.1	-0.5	-2.6

Table 5: Cross-utterance accuracy for various combinations of same and different talker and test word, along with comparisons between them (Δ). Bold entries are statistically significant.

Note that a number of the cross-utterance accuracies in Table 3 are higher than the within-utterance cross-validation results in Table 2. This is rather surprising, as the cross-validation results only require generalizing across noise instances, whereas the cross-utterance results also require generalizing across utterances. Table 4 shows that this is mainly due to the increased number of examples available to train the cross-utterance classifiers as compared to the cross-validation classifiers. Specifically, it shows the accuracy of cross-utterance classifiers trained on the same number of examples as the cross-validation classifiers for each utterance. The number of training points, averaged across utterances, is shown in the N_{tr} column. In the limited data condition, cross-validation models generally outperform cross-utterance models.

Table 5 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. The top half of the table shows that these models are able to predict the intelligibility of mixtures significantly better than chance for all words in all combinations of same and different talker and word. 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, we test each of the classifiers trained in the same word condition 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 results show that the classifiers can achieve an accuracy of 57–64% in these conditions, all of which are significant. This is due to the fact that all of the utterances share the same general structure and pacing, i.e., the important regions for distinguishing between these words are all approximately aligned. Aligning the different words to a reference has little effect on prediction accuracy.

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. It shows that the intelligibilities of mixtures involving the words *ada*, *ata*, and *ava* are predicted significantly better than the pooled cross-word results. This means that the classifiers for those words capture a significant amount of phoneme-specific information beyond the general structure shared by all of the words. The bottom half of the table also shows that *afa* is significantly better able to generalize across utterances from the same talker than from different talkers, indicating that those classifiers are able to capture a significant amount of talker-specific information.

6. Conclusion

We have shown that the intelligibility prediction framework of [10] is able to generalize not only to novel noise instances, but also to novel instances of the same word from the same talker and from different talkers, and to novel phonemes in *aCa* nonsense words. These abilities are the necessary first steps toward creating a classifier-based intelligibility predictor able to generalize to arbitrary new mixtures from a finite amount of human-generated training data. While 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 completely different words and to the same words in different phonetic contexts.

7. References

[1] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acous. Soc. Am.*, vol. 88, pp. 1725–1736, Oct. 1990.

[2] J. I. Alcántara, B. C. J. Moore, V. Kühnel, and S. Launer, "Evaluation of the noise reduction system in a commercial digital hearing aid," *Int J Audiol*, vol. 42, pp. 34–42, Jan. 2003.

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

[4] R. Lippmann, "Speech recognition by machines and humans," *Speech Comm.*, vol. 22, pp. 1–15, Jul. 1997.

[5] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Comm.*, vol. 49, pp. 336–347, May 2007.

[6] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *J. Acous. Soc. Am.*, vol. 131, pp. EL256–EL261, Mar. 2012.

[7] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acous. Soc. Am.*, vol. 120, pp. 4007–4018, Dec. 2006.

[8] N. Li and P. C. Loizou, "Factors influencing glimpsing of speech in noise," *J. Acous. Soc. Am.*, vol. 122, pp. 1165–1172, Aug. 2007.

[9] F. Apoux and E. W. Healy, "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," *Hearing Research*, vol. 255, pp. 99–108, Sep. 2009.

[10] M. I. Mandel, "Learning an intelligibility map of individual utterances," in *Proc. IEEE WASPAA*, 2013.

[11] E. W. Healy, S. E. Yoho, and F. Apoux, "Band importance for sentences and words reexamined," *J. Acous. Soc. Am.*, vol. 133, pp. 463–473, Jan. 2013.

[12] 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, pp. 3387–3405, May 2009.

[13] 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, pp. 2599–2610, Apr. 2010.

[14] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, pp. 2261–2271, Aug. 2001.

[15] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, "Using auditory classification images for the identification of fine acoustic cues used in speech perception," *Frontiers in Human Neuroscience*, vol. 7, 2013.

[16] R. V. Shannon, A. Jensvold, M. Padilla, M. E. Robert, and X. Wang, "Consonant recordings for speech testing," *J. Acous. Soc. Am.*, vol. 106, no. 6, pp. L71+, 1999.

[17] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, Aug. 1990.

[18] ANSI, "American national standard methods for manual pure-tone threshold audiometry," Tech. Rep. S3.21 (R2009), Acoustical Society of America, 2004.

[19] ANSI, "American national standard specification for audiometers," Tech. Rep. S3.6, Acoustical Society of America, 2010.

[20] L. Calandruccio and K. A. Doherty, "Spectral weighting strategies for sentences measured by a correlational method," *J. Acous. Soc. Am.*, vol. 121, no. 6, pp. 3827–3836, 2007.

[21] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. ECML*, pp. 39–50, Springer, 2004.

[22] D. Ellis, "Dynamic time warp (DTW) in matlab." Web resource, available: `http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/`, 2003.