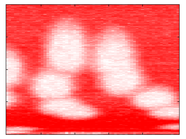


Learning an intelligibility map of individual utterances

Michael I. Mandel · Computer Science & Engineering · mandelm@cse.osu.edu

The Ohio State University

Problem

Predicting the intelligibility of speech in noise is difficult. Most current algorithms aim to be correct on average across many recordings, ignoring important variations caused by spectro-temporal characteristics of the speech and noise. This poster describes a listening test paradigm and associated analysis techniques that can predict the intelligibility of a specific recording of a word in the presence of a specific unseen noise instance, capturing such characteristics.

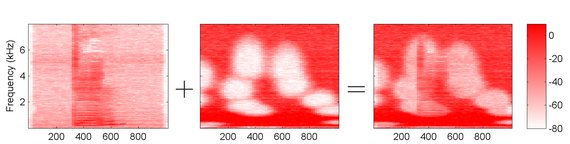
Contributions

- Designed noise to allow "glimpses" (Cooke, 2006) of random spectrogram regions
- Formulated descriptive model that identifies spectral regions important to an utterance's intelligibility by measuring intelligibility of mixtures of it with different noise instances
- Formulated predictive model that can generalize to unseen noise instances by treating intelligibility prediction as a classification problem

Listening test

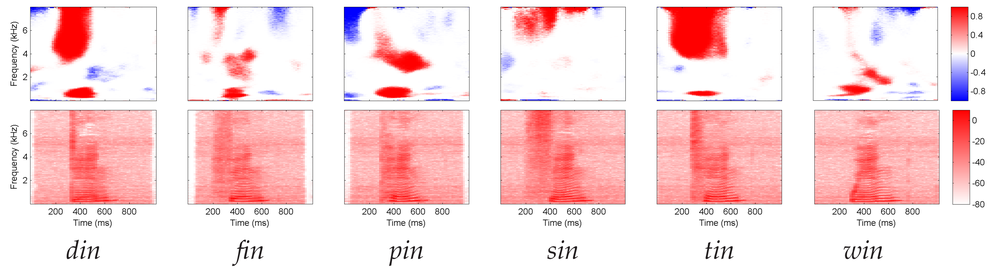
- Six-way forced choice paradigm
- Speech material: Consonants in *Cin*, Hillenbrand vowels in *hVd*
  - Consonants: *din*, *fin*, *pin*, *sin*, *tin*, *win*
  - Vowels: *had*, *head*, *heed*, *hid*, *hood*, *who'd*
- Noise: speech shaped noise at SNR -30 dB with "bubbles" or "glimpses" cut out
- Listeners: one expert (author), 74 workers on Amazon's mechanical Turk
- 199 mixtures for each target recording
- Set bubbles per second to attain $\sim 50\%$ correct
- Some mixtures heard by multiple "Turkers"

Example mixture



Descriptive model

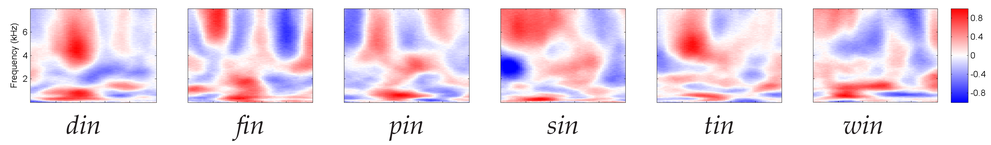
- **Goal**: Identify spectrogram points that contribute to mixture intelligibility for one word
- **Method**: Identify spectrogram points that were glimpsed more often in intelligible mixtures
- **Specifics**: Pearson's chi-squared test of independence for each spectrogram point



Top: Spectrogram regions for that increase (red) and decrease (blue) intelligibility more than would be expected by chance. Bottom: corresponding spectrograms of clean speech.

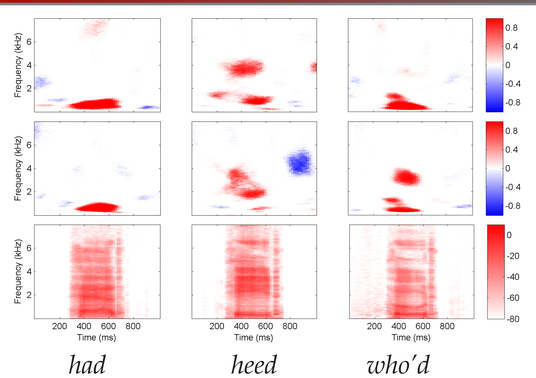
Predictive model

- **Goal**: Predict intelligibility of known utterance with new noise instance
- **Method**: Support Vector Machine (SVM) trained to predict intelligibility of mixtures for each utterance from Principal Components Analysis (PCA) of noise "bubbles"
- **Specifics**: Nested cross-validation to choose PCA dimension, measure generalization



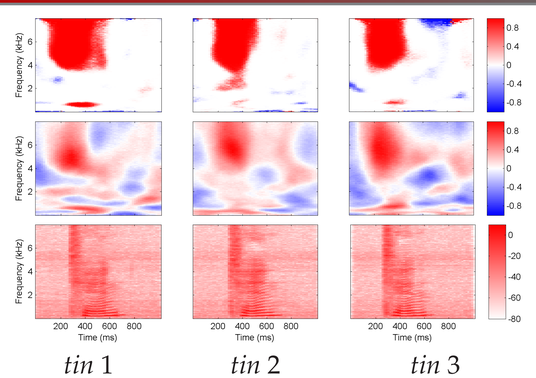
Visualization of spectral regions that increase (red) and decrease (blue) intelligibility prediction of SVM.

Naïve vs expert listeners



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

Repeated utterances



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).

Measuring generalization

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

Conclusions

- Descriptive model identifies regions that agree with the acoustic phonetics literature
- Predictive model can generalize to unseen noise instances
- Predictive model also agreed with descriptive model and acoustic phonetics
- Mechanical Turk listeners give similar results to expert listener (with repeated listenings)
- Descriptive and predictive models are consistent across repeated utterances

Future Directions

- Generalize predictions to unseen utterances and unseen talkers
- Compare intelligibility maps of normal hearing and hearing impaired listeners
- Compare intelligibility maps of humans and automatic speech recognizers
- Use intelligibility maps to evaluate noise suppression algorithms

References

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

Li, F., A. Menon, and J. B. Allen (2010). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech". In: *J. Acous. Soc. Am.* 127.4, pp. 2599–2610.

Mandel, M. I. (2013). "Learning an intelligibility map of individual utterances". In: *Proc. IEEE WASPAA*. To appear.