Relating automatic vowel space estimates to talker intelligibility

Yi Luan[1]*, Richard Wright*[1]*, Mari Ostendorf*[1]*, Gina-Anne Levow*[1]

[1]University of Washington

{luanyi, rawright, ostendor, levow}@uw.edu

Abstract

Differences in pronunciation have been shown to underlie significant talker-dependent intelligibility differences. There are several dimensions of variability that are correlated with talker intelligibility including pitch range, vowel-space expansion, and rhythmic patterns. Prior work has shown that some of the better predictors of individual intelligibility are based on the talker's F1 by F2 vowel space, but findings are based on hand-corrected measurements on carefully balanced sets of vowels, making large scale analysis impractical. This paper proposes a novel method for automatic estimation of a talker's vowel space using sparse expanded vowel space representations, including an approximate convex hull sampling, which are projected to a low dimensional space for intelligibility scoring. Both supervised and unsupervised mappings are used to generate an intelligibility score. Automatic intelligibility rankings are assessed in terms of correlation with an intelligibility score based on human transcription accuracy. We find that including a larger sample of vowels (beyond point vowels) leads to improved performance, obtaining correlations of roughly 0.6 for this feature alone, which is a strong result given that there are other factors that may also contribute to a talker's intelligibility in addition to a talker's vowel space area.

Index Terms: phonetic analysis, intelligibility, automatic estimation, convex hull

1. Introduction

Talker-dependent differences in pronunciation have been shown to underlie significant intelligibility differences. There are several dimensions of variability that are correlated with talker intelligibility including pitch range, vowel-space expansion measured as the distance to the vowel space center, and rhythmic patterns related to lexical stress [1] [2] [3]. Other factors that are easily extracted, such as speech rate, mean pitch, or talker gender are unreliable as predictors of intelligibility [1] [2]. In a recent study, McCloy et al. [2] found that the convex hull of a talker's first formant (F1) by second formant (F2) vowel space is a robust predictor of intelligibility differences in a gender-balanced corpus of 20 talkers. The convex hull is also of interest because the expansion or contraction of a talker's vowel space is indicative of within-talker variability related to clear speech adjustments [3]. However, as the McCloy et al. study illustrates, vowel space expansion and contraction estimates have typically involved hand measurements and well balanced sets of vowels, making them impractical for large scale or automated data collection. In this study we estimate the convex hull based on a relatively sparse sampling of a vowel space using automatically extracted vowel formants on a corpus that has been forced aligned. To test our estimated vowel-space measures, we used them to predict the intelligibility of the set of talkers in the corpus used in McCloy et al.'s intelligibility study [4].

The first phase of this study is establishing an efficient and reliable method for automatically estimating a talker's vowel space. First, the F1 and F2 of each speaker's vowel space are automatically extracted from a corpus of read-sentences that had previously been automatically aligned but not hand corrected. The F1 and F2 measures are then normalized using a vocal tract length warping and plotted. A novel low-dimensional convex hull estimation method is proposed and implemented, with a sparse convex hull being reconstructed using the low dimensional data. The resulting sparse convex hull is tested against the original convex hull, resulting in a very close approximation. The relative contribution of the vowel categories was investigated and it was found that point vowels are the main contributors to the convex hull. The second phase of this study is evaluating utility of different transformations of the vowel space representation for predicting intelligibility, evaluating both supervised and unsupervised methods in terms of the correlation of ranking with respect to human scores.

The reminder of the paper is organized as follows. In Section 2, we give a brief review on McCloy et al.'s corpus and the definition of intelligibility scores. In Section 3, we propose the method to capture the convex hull of the vowel space using low dimensional features (sparse convex hull). The result of average contribution of each vowel to the sparse convex hull is also introduced. Then the proposed PCA and GLM methods to model intelligibility are introduced in Section 4. The results of experiments are described in Section 5. Finally, Section 6 draws overall conclusions and describes possible future work.

2. Speech materials and intelligibility scores

2.1. Speech material

McCloy et al.'s corpus is made up of recordings of 20 talkers: 5 male and 5 female talkers each from two dialect regions (the West and Northern Cities) [6]. Each talker read the same subset of 180 of the IEEE sentences [5] using a uniform reading style [4]. The speech signal was force aligned using the reading material for the sentences with the CMU dictionary pronunciations (corrected for gross dialect differences) [8]. In our test we use only the automatic alignments rather than the subset of hand-corrected alignments used in McCloy et al.'s [4] study. This gives us 180 read-sentences per talker as a basis for estimating vowel space expansion.

2.2. Intelligibility scores

The intelligibility scores for each talker were based on McCloy et al.'s data. In their study, the sentences were presented in quiet and in speech shaped noise (+6 and +2 dB SNR) to 28 listeners, 15 from the West and 13 from the Northern Cities. The masker in the noise conditions was Gaussian noise filtered to match the long-term spectral average of the corpus. To ensure target au-

dibility, the level of the speech was held constant at 68 dB SPL (dB RMS in a 6 cc coupler) and different levels of masker noise were digitally added to the speech to achieve the desired SNRs. The listeners repeated the sentences that they heard and were instructed to guess when they were unsure. Their responses were recorded and scored as 0-5 keywords correct. To avoid repetition effects, each listener heard each sentence only once. For our intelligibility ranking we use the +2 dB SNR condition, which was the only level that avoided ceiling effects for all talkers. This gives us a speaker intelligibility ranking that is based on 28 data points (utterance scores) per talker.

3. Vowel space characterization

A goal of this work is to develop an automatic approximation of the careful acoustic measurements used in the McCloy et al. study that extends to unrestricted speech. Key differences in the methods used here include:

- Heuristics are developed to automate selection of the measurement point in the syllable, and outliers are excluded to minimize the impact of formant extraction errors.
- McCloy et al. measured only the full vowels in target words used in the intellibility study. Here, measurements are included for all full vowels in lexically stressed syllables, as well as contrasting measurements for unstressed vowels. In addition, all 180 sentences recorded by the speaker were used here, whereas only a subset were used by McCloy et al.
- Speaker variation associated with vocal tract length differences is accounted for here using standard full spectrum normalization methods from speech recognition.
- For purposes of developing a fixed-dimension representation, a sparse sampling of the space is introduced.

Further details are described in this section. Experiments assess the utility of including different subsets of vowels in the representation.

3.1. F1 F2 extraction

The vowel space characterization was based on a measurement of the first and second formants (F1 and F2) of all the vowels in each sentence. Energy, F1 and F2 contours were automatically extracted at a 10ms frame rate using Praat [7], specifically using the Linear Predictive Coding (LPC) based formant tracker with a window length of 25 ms and a dynamic range of 30 dB. For vowels that were 80 ms long or greater, formant values were measured intensity peak of the vowel, which was determined using a second-order polynomial regression fit to the time sequence of (linear) intensity values. When there was not a clear peak (correlation curvature $a > -500$) or for vowels that were less than 80 ms, the measure was made at the temporal midpoint to extract F1 and F2. Based on pilot data results, we used a setting of 5 formants in the 0-5500 Hz range for females, and for males either a 5 or a 6 formant setting was used based on the one that resulted in fewer outliers in a two-sigma ellipse. A sigma ellipse was calculated for each vowel for each talker and outliers were discarded. The function to calculate the sigma ellipse is as follows:

$$\frac{(\cos(\alpha)(F1-\bar{F1})+\sin(\alpha)(F2-\bar{F2}))^2}{a^2}+ \\ \frac{(\sin(\alpha)(F1-\bar{F1})+\cos(\alpha)(F2-\bar{F2}))^2}{b^2} \leq 1 \quad (1)$$

where $a = s\sqrt{\lambda_2}$ and $b = s\sqrt{\lambda_1}$. λ_1 and λ_2 are the smallest and largest eigenvectors of the feature covariance matrix, and s is the number of standard deviations ($s = 2$ here). To reduce the impact of formant tracking errors, the points outside of the 2-sigma ellipse are discarded (corresponding to roughly 12% of the vowels). Eliminating data is not a big problem in this application, though it may be problematic for working wiht shorter speech segments.

Table 1: Acoustic GMM model condition

GMM mixture	256
Warp Scales α	[0.90, 1.10] every 0.01 step
Feature	MFCC_E_D_A
Sample Rate	41000
Warp low cut frequency	100
Warp high cut frequency	16000

In order to reduce variation due to speaker differences, vocal tract length normalization is implemented before the vowel space analysis. In order to maintain the interpretation in formant space, we chose to use frequency warping (i.e. F1'=αF1, F2'=αF2) rather than mean normalization. The warping factor $\hat{\alpha}$ is estimated using a popular technique in speech recognition that chooses the warping factor from a fixed set of possibilities determined by maximizing the likelihood of all utterances of a speaker with respect to a given Gaussian Mixture Model (GMM) built using data from all the speakers [9]. The conditions for training the GMM model are given in Table 1; training data includes all utterances (unwarped) from all speakers in McCloy's corpus. To determine the warping factor for a speaker, the set of utterances from the speaker are processed for a range of warping factors $\alpha \in [0.9, 1.1]$ with step size .01. Let $\boldsymbol{X}_i^\alpha$ denote the cepstrum domain (MFCC) observation vectors for a set of utterances from speaker i warped by α. The optimal warping factor for speaker i is given by:

$$\hat{\alpha}_i = \underset{\alpha}{\operatorname{argmax}}\, Pr(\boldsymbol{X}_i^\alpha | G) \quad (2)$$

where G denotes the common GMM model. In pilot experiments with the TIMIT corpus, we found that cross-speaker variability (as measured by scatter matrices of the vowel means) was minimized using this warping method compared to the parametric approach in [10], and that the method is effective for normalizing across genders.

3.2. Convex hull estimation

Let $\boldsymbol{O}_{vk}^s$ be the two-dimensional normalized (F1, F2) measurement for the k-th instance of vowel v from speaker s. We calculate the mean from all the instances for vowel v:

$$\mu_v = \frac{1}{\sum_s \sum_k 1} \sum_s \sum_k \boldsymbol{O}_{vk}^s \quad (3)$$

to obtain the universal mean for all vowels:

$$\boldsymbol{c} = \frac{1}{V} \sum_v \mu_v \quad (4)$$

A fixed-dimension sparse representation of the size of the vowel space of speaker s is constructed by selecting two instances $\boldsymbol{O}_{vk}^s$ of each vowel v distant from $\boldsymbol{c}$ in different directions (F1 vs. F2) towards the convex hull. Specifically for F1, choose the point that is maximally different from $\boldsymbol{c}$ in F1, with ties broken by

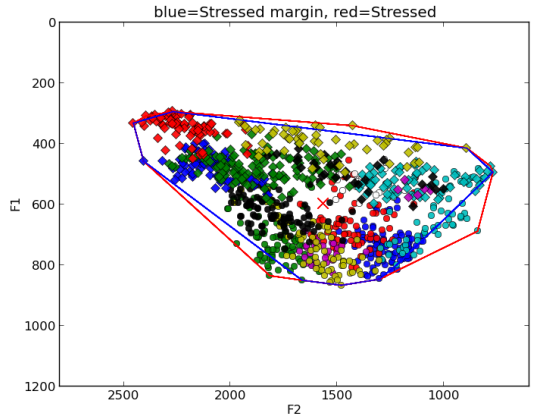


Figure 1: *Worst case example of sparse convex hull*

Table 2: Most and least frequent full vowel categories

Frequency $= 0$	AH, EH, ER, EY, IH, OW, OY, UH
Frequency $\in [0.1, 0.45]$	AO, AW
Frequency $\in [0.6, 1]$	AA, AE, AY, IY, UW

maximizing the total squared distance. The F2 point is chosen similarly.

This sparse representation tends to include a good sampling of the points that are included in the convex hull of the full set of observations, such that the convex hull of the sparse representation is typically quite close to that of the full set. Figure 1 gives the worst case of the convex hull reconstruction for this data set. The worst case misses a few corners of the convex hull, but the overall volume is not changed substantially.

The sparse vowel space representation was used to find convex hulls for each of the 20 speakers in the corpus, from which we calculated the relative frequency that each of the 2V sample points is used in the hull. As expected, only a subset of the vowels are frequently used, specifically the set of point vowels (AA, AE, IY, UW) and the diphthong (AY) were used. Table 2 groups the full vowels according to their relative frequency of usage in defining the sparse convex hull.

Unstressed vowels are not often used in vowel space characterization because they can be highly variable. However, we hypothesized that differences in the stressed and unstressed vowel space sizes might be related to intelligibility, so samples of these vowels were also included in an additional experiment.

3.3. PCA Projection

The sparse representation is still high dimensional, depending on the number of vowels included, so it is projected to a lower dimension using principal components analysis (PCA). The vectors for speaker s are stacked into matrix $\boldsymbol{X} \in \mathbb{R}^{2V \times S}$, where S is the total number of speakers and V is the total categories of vowels. Then the covariance matrix $\boldsymbol{M}$ is defined as

$$\boldsymbol{M} = \frac{1}{S}(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top \tag{5}$$

where $\bar{\boldsymbol{X}}$ is the vector mean. Let $\boldsymbol{U} = [\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top, ...\boldsymbol{u}_i^\top]$, where $\boldsymbol{u}_1$ to $\boldsymbol{u}_i$ are the eigenvectors that give the first i largest eigenvalues of $\boldsymbol{M}$. The projected low dimension matrix $\boldsymbol{W}$ is obtained by

$$\boldsymbol{W} = \boldsymbol{U}^\top(\boldsymbol{X} - \bar{\boldsymbol{X}}) \tag{6}$$

When working with a small number of sentences, some of the speakers may lack examples of some vowels, in which case the feature vector will not be complete. In this case, we use Probabilistic Principle Component Analysis (PPCA) [11] to fill in the blanks in the feature vector.

4. Modeling intelligibility

Two approaches are investigated for automatically predicting intelligibility: i) an unsupervised approach that approximates the volume of the vowel space in terms of the magnitude of the projected vector, and ii) a supervised learning approach using a generalized linear model (GLM) on the projected space, as described below.

Unsupervised approach. The unsupervised approach uses the magnitude of the PCA projection as the machine score for intelligibility. This is a rough measure of estimating the volume of the convex hull, motivated by the results of McCloy et al. Specifically, given the D-dimensional projected vector $\boldsymbol{w}_s$ for speaker s, the intelligibility score vector $\boldsymbol{g}_s$ is obtained by

$$\boldsymbol{g}_s = \boldsymbol{w}_s^\top \boldsymbol{w}_s \tag{7}$$

Supervised approach. Since $\boldsymbol{g}_s$ is only a rough estimate of the vowel space, we also investigated the use of an automatically learned regression function mapping from the projected space to intelligibility. Given that there are very few speaker samples in this data set, we used a very simple GLM regression mapping trained using leave-one-out cross validation. As before, the sparse speaker vowel samples are projected to the D-dimensional vector $\boldsymbol{w}_s$. Let $\boldsymbol{y}_s$ be the intelligibility score for speaker s and $\boldsymbol{b}$ be the parameter vector. The GLM learns a mapping

$$F^{-1}(\boldsymbol{y}_s) = \boldsymbol{w}_s^\top \boldsymbol{b} \tag{8}$$

where F^{-1} is the inverse Continuous Density Function of a normal distribution, chosen since $\boldsymbol{y}_s \in [0, 1]$.

5. Experimental Results

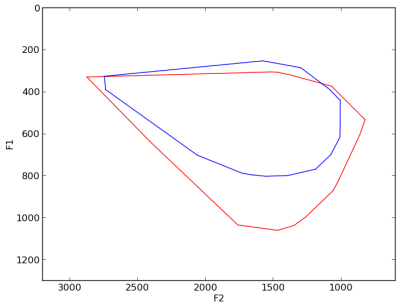
We conducted two sets of experiments to answer the following questions:

- Which vowels are useful for characterizing vowel space expansion for predicting intelligibility?
- How much can be gained by directly learning a function to predict intelligibility over the simple measure of vowel space volume?

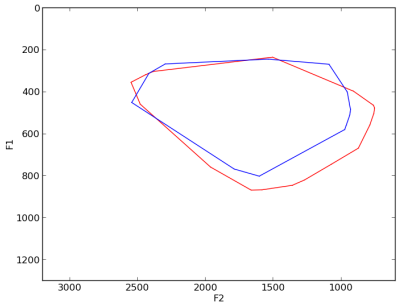
The correlation of the machine score and human score is calculated by Spearman's rank correlation, which uses the rank of the scores instead of the absolute score to do correlation.

We expected that the sparse convex hull vowels would be most useful, based on previous results. In addition, we hypothesized that the unstressed vowels would only be useful if treated differently than the stressed vowels, assuming that greater differences between stressed and unstressed vowels would lead to greater intelligibility. As shown in figure 2, comparing the most and least intelligible speakers shows a bigger difference in the convex hulls of stressed vs. unstressed vowels for more intelligible speakers.

The experiments on different vowel measurements compared low-dimensional projections of three different sparse



(a) Stressed convex hull(red) v.s. unstressed convex hull (blue) of the most intelligible speaker



(b) Stressed convex hull(red) v.s. unstressed convex hull (blue) of the least intelligible speaker

Figure 2: Stressed and unstressed convex hull of the most and least intelligible speaker.

Table 3: Correlation between unsupervised machine and human intelligibility scores for different vowel space representations

Vowel Subset	PCA dim	corr
Stressed	3	0.524
Hull	3	0.264
All	2	0.220

Table 4: Correlation between GLM and human intelligibility scores for different vowel space representatIons

Vowel Subset	PCA dim	corr
Stressed	6	0.590
Hull	6	0.324
All	4	0.719

samplings: the set of 2 samples each from 15 full vowels measured in stressed syllables (60 dimensions), the set of 2 samples each from 5 frequent convex hull vowels (20 dimensions), and 2 samples each from the combination of the 15 stressed vowels and the 6 full vowels on the unstressed convex hull (84 dimensions). Table 3 shows the results with for the best case dimension using the unsupervised volume estimate. Using the full set of stressed vowels (with sparse measurements) gives better results than the vowels on the sparse convex hull. While the interior vowels are not used in finding the full vowel space, including them may provide a more robust estimate of vowel space expansion when reduced to a lower dimension. Adding the unstressed vowels hurts performance, as expected, since in this case they are treated the same as stressed vowels.

Table 4 gives the results for the same starting set of features but using a GLM, again optimizing for the best reduced dimension. For the stressed vowel features, there is only a small gain from supervised learning with double the dimensions, so we conclude that the projected vector magnitude is a good approximation of volume. For the unstressed vowels, the use of only a 2-dimensional projection leads to improved prediction performance, presumably because the difference between stressed and unstressed vowels can be captured. This leads to the best case correlaton of 0.72. The associated correlation scatter plot is given in Fig. 3.

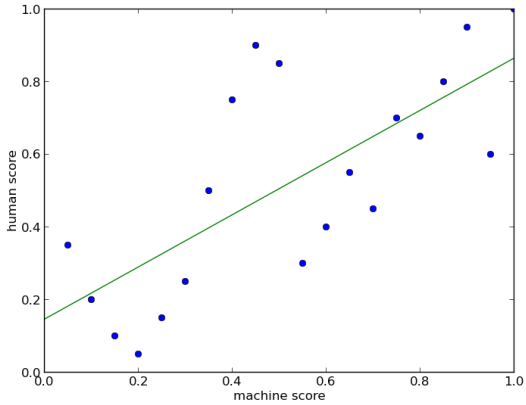


Figure 3: *Correlation plot of all speakers*

6. Conclusions

In this paper, we propose a novel automatic estimation of talkers' vowel spaces using sparse convex hull representations from automatic formant extraction. The resulting estimates were used to realize a fully automatic estimation of talker-intelligibility. Both supervised and unsupervised mappings are used to generate intelligibility scores. The automatic intelligibility rankings are assessed in terms of correlations with an intelligibility score derived from human recognition accuracy. We find that including a larger sample of stressed vowels (beyond point vowels) leads to improved performance, obtaining a correlation of roughly 0.6 for this feature alone. We obtain a better result using both stressed and unstressed vowel convex hull representations with a correlation of 0.72. We consider these correlations to be strong results given that there are several other factors that may also contribute to a talker's intelligibility in addition to a talker's vowel space area.

7. Acknowledgements

This research is supported by the US National Science Foundation, IIS: #1351034. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

8. References

[1] Bradlow, A. R., Torretta, B. M., Pisoni, D. B., 1996. Intelligibility of normal speech I: Global and ne-grained acoustic-phonetic talker characteristics. Speech Communication 20, 255272.

[2] McCloy, D. R., Wright, R. A., & Souza, P. E. (2014). Models of intelligibility variation: Prosodic and vowel-space predictors. Proceedings of Meetings on Acoustics, vol 18.

[3] Picheny M. A., Durlach N. I., & Braida L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. J. Speech Hear. Res., 28(1), 96-103.

[4] McCloy, D. R., Souza, P. E., Wright, R. A., Haywood, J., Gehani, N., & Rudolph, S. (2013). The PN/NC corpus (Version 1.0). Seattle: University of Washington. Retrieved from http://depts.washington.edu/phonlab/resources/pnnc/

[5] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. IEEE Transactions on Audio and Electroacoustics, 17, 225246. DOI: 10.1109/TAU.1969.1162058

[6] Labov, W., Ash, S. & Boberg, C. (2006). The Atlas of North American English. New York: Mouton de Gruyter.

[7] Boersma, P. & Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.66, retrieved 9 March 2014 from http://www.praat.org/

[8] Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. Proceedings of Acoustics '08

[9] Lee, L. and Rose, R. Speaker normalization using efficient frequency warping procedures. *ICASSP, 1996*

[10] Eide, E., and Gish, H., "A parametric approach to vocal tract length normalization," *Proc. ICASSP,* 1996.

[11] Verbeek, J., Vlassis, N., and Krose, B. 2002. Coordinating principal component analyzers, Articial Neural NetworksICANN 2002, pp. 140140, 2002.