INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Comparison of Sentence-level Speech Intelligibility Metrics

Alexander Kain[1], *Max Del Giudice, Kris Tjaden*[2]

[1]Center for Spoken Language Understanding, Oregon Health & Science University
Portland, OR, USA
[2]University at Buffalo, Buffalo, NY, USA

kaina@ohsu.edu, max.delgiudice@urbanairship.com, tjaden@buffalo.edu

Abstract

We examine existing and novel automatically-derived acoustic metrics that are predictive of speech intelligibility. We hypothesize that the degree of variability in feature space is correlated with the extent of a speaker's phonemic inventory, their degree of articulatory displacements, and thus with their degree of perceived speech intelligibility. We begin by using fully-automatic F1/F2 formant frequency trajectories for both vowel space area calculation and as input to a proposed class-separability metric. We then switch to representing vowels by means of short-term spectral features, and measure vowel separability in that space. Finally, we consider the case where phonetic labeling is unavailable; here we calculate short-term spectral features for the entire speech utterance and then estimate their entropy based on the length of a minimum spanning tree. In an alternative approach, we propose to first segment the speech signal using a hidden Markov model, and then calculate spectral feature separability based on the automatically-derived classes. We apply all approaches to a database with healthy controls as well as speakers with mild dysarthria, and report the resulting coefficients of determination.

Index Terms: vowel space area, intelligibility

1. Introduction

We aim to develop automatically derived acoustic metrics that predict sentence-level speech intelligibility (in contrast to single word intelligibility). We assume that larger variations in feature space are associated with a larger phonemic inventory, greater articulatory displacements, and thus increased speech intelligibility. In this paper, we examine three metrics that measure this feature space variation, based on the following:

Vowel-based formant frequencies These features are commonly used in the literature, and require both phonetic labeling (to identify the location and extent of the vowels of interest) and formant tracking (to identify the trajectories of the spectral peaks of the spectrum, caused by the acoustic resonance of the human vocal tract). While automatic methods exist for both of these requirements, a human expert usually performs corrections to ensure the highest possible accuracy. Using these features, a vowel space area (VSA) in Hz^2 can be computed by averaging features over all instances of certain vowels (e. g. the peripheral vowels), and then creating a polygon (commonly a quadrilateral) whose vertices are defined by those average feature vectors. In an alternate approach, we propose a measure of class separability based on pair-wise distances, and apply it to the formant frequency trajectories of distinct vowels.

Vowel-based spectra We propose to replace formant frequencies by short-term spectral envelope features. We then calculate the class-separability using those features. Our hypothesis is that the this measure is similar in predictive performance to one based on formant frequency trajectories, but with the advantage of avoiding formant tracking. Phonetic labeling is still required.

Unsupervised spectra We propose to consider the entire utterance, as opposed to merely vowel centers. This alleviates the requirement for phonetic labeling, and it has the potential to improve prediction, since the entire signal is available during analysis. In a first approach, we apply a previously-published entropy measure. In a second approach, we propose to cluster the signal using a hidden Markov model, and to calculate the class-separability of the resulting segmentation.

We will examine the correlations between these metrics and sentence-level speech intelligibility, using both typical as well as speakers diagnosed with Multiple Sclerosis and Parkinson's disease.

2. Background

A longstanding challenge for speech research has been to develop acoustic metrics predictive of intelligibility for normal speakers as well as for clinical populations, such as individuals with dysarthria. Efforts to date have largely focused on metrics of vowel segmental integrity derived from the first (F1) and second (F2) formant frequencies [1, 2, 3, 4, 5, 6, 7]. Vowel space area (VSA) refers to the two-dimensional quadrilateral space formed by F1 and F2 measured at a static point in time for the four English peripheral vowels (/i/, /æ/, /a/, and /u/) or non-peripheral vowels (/ɪ/, /ʊ/, /ʌ/, and /ɛ/) [5]; for the purposes of this study, we will focus on peripheral-VSA only.

Although the strength of the relationship varies substantially across studies, larger VSAs, which indicate increased articulatory-acoustic distinctiveness, are associated with increased intelligibility [6, 7, 8]. Procedures for optimizing the strength of association between perceived speech adequacy and vowel metrics requiring static measures of F1 and F2 have recently been proposed [7]. However, VSA and related metrics for quantifying acoustic distinctiveness among vowel categories have several limitations. First, vowel segments require labeling. Automated labeling routines may perform reasonably well for normal speech, but are typically unsatisfactory for disordered speech signals. Second, any computer-generated formant tracking errors also must be manually corrected. An alternative metric of segmental articulatory-acoustic distinctiveness for normal

This material is based upon work supported by the NIH under Grant R01DC004689.

http://dx.doi.org/10.21437/Interspeech.2017-567

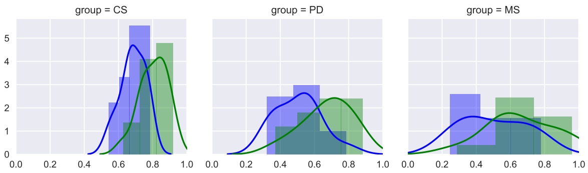


Figure 1: *Histograms and kernel density estimations of scaled (blue) and transcription intelligibility (green) responses, for CS, MS and PD groups.*

and clinical populations that does not rely on phonetic segmentation or manual correction of formant tracking errors and that also correlates with intelligibility is therefore desirable [9, 10].

3. Corpus

The corpus used in this study was comprised of sentence productions of healthy controls (CS) as well as speakers with Multiple Sclerosis (MS) or Parkinson's disease (PD). Prior publications [5, 11, 12] report VSAs and intelligibility for these speakers and speech materials; it was found that speakers with MS or PD had mostly mild to moderate dysarthria, as indicated by the Sentence Intelligibility Test (SIT) [13] and auditory-perceptual judgments of speech-language pathologists. The SIT is a widely used clinical tool for quantifying sentence-level intelligibility in dysarthria. All 15 control, 11 MS, and 13 PD speakers read the same 24 Harvard sentences [14] in their habitual or typical manner, for a total of 39×24=936 sentences. The acoustic signal was sampled at 22 kHz and down-sampled to 8 kHz for the current study.

Harvard sentences included three to five occurrences of the four peripheral vowels. Vowel segments were manually identified and computer-generated formant tracking errors were corrected by trained personnel [5]. F1 and F2 values were extracted from the temporal midpoint of vowels and formant values were averaged across all occurrences of a given vowel. For each speaker, a (peripheral) VSA was calculated using Heron's formula, as described in our prior publication [5].

In perceptual tests, 50 listeners judged the intelligibility of the sentences using a continuous Visual Analog Scale, for the remainder of the paper referred to as scaled intelligibility (SI), with possible scale values ranging from 0.0 (completely understandable) to 1.0 (can't understand anything); for the purposes of the paper we transformed SI=1-SI [11]. An additional 50 listener orthographically transcribed the sentences, resulting in percentage of words correctly transcribed, hereafter referred to as transcription intelligibility (TI) [12]. The sentences were amplitude-normalized and mixed with multi-talker babble prior to presentation to listeners to prevent ceiling effects.

Fig. 1 shows the distribution of SI and TI intelligibility, faceted by group. Note that SI and TI distributions are closely aligned; in fact SI is predictive of TI at r^2=0.82. Furthermore, note that SI and TI intelligibility spanned a relatively restricted range for all speakers. We calculated the coefficient of determination for VSA→SI at r^2=0.03 (p=0.27), and for VSA→TI r^2=0.08 (p=0.08). We hypothesize that the fact that SI and TI did not differentiate the three speaker groups, or result in higher r^2 values, reflects the mild dysarthria for speakers with MS or PD, as well as the limited range of intelligibility scores for all speakers.

4. Vowel Formant Frequency Metrics

In this section, we use fully-automatic formant frequency trajectories as features, and measures will require knowledge of vowel regions. We used Praat [15] to automatically generate F1 and F2 formant frequency trajectories for the entire corpus. For any instance of a vowel of interest, we restricted analysis to the center 34% of the segment.

4.1. Vowel space area

In this subsection we aim to reproduce our former study, except with fully automatic formant frequency tracks for calculation of the VSA, instead of manually corrected formant tracks [5]. We first computed the average formant frequency vector for each vowel instance. Using these data, we computed the global average formant vectors for each vowel, separately for each speaker. Then we considered the global average vowel vectors as vertices of a polygon, and calculated the area via the Shoelace formula; we will use the term AVSA for the automatically-derived VSA.

We predicted AVSA→VSA and found r^2= 0.86; in other words, using automatically-estimated formant frequency trajectories give a good first estimate of VSA, as compared to manually-corrected ones.

We then predicted AVSA→SI and AVSA→TI, and found r^2=0.04 (p=0.21) and 0.11 (p=0.04), respectively; this is comparable to the results in Section 3. When performing the analysis separately for CS vs MS/PD speakers, we found, for the former, r^2=0.10 for CS and 0.07 for MS/PD; for the latter, r^2=0.23 for CS and 0.14 for MS/PD. We speculate that the group differences in the coefficients' values may at least in part be caused by formant tracking performing worse for disordered speech.

4.2. Pairwise-distance based separability

Another way of modeling variation in F1/F2 formant space is by considering the separability between classes. Recently, Lansford and Liss [6] used a dispersion measure based on the Euclidian distance between pairs of features. However, it is possible to simultaneously observe a large dispersion and a large overlap. To address this, we define a separability measure S that is also based on average pair-wise distances between data points, stored as N rows of D-dimensional vectors in matrix $\mathbf{X}_{N\times D} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top$ with $\mathbf{x} = [x_1, x_2, \ldots, x_D]^\top$, but which considers both between- and within-class distances. Specifically, given C classes, we calculate the separability S by forming the ratio

$$S = \frac{B}{W} \tag{1}$$

where B is the average pair-wise Euclidian distances of data *between* classes

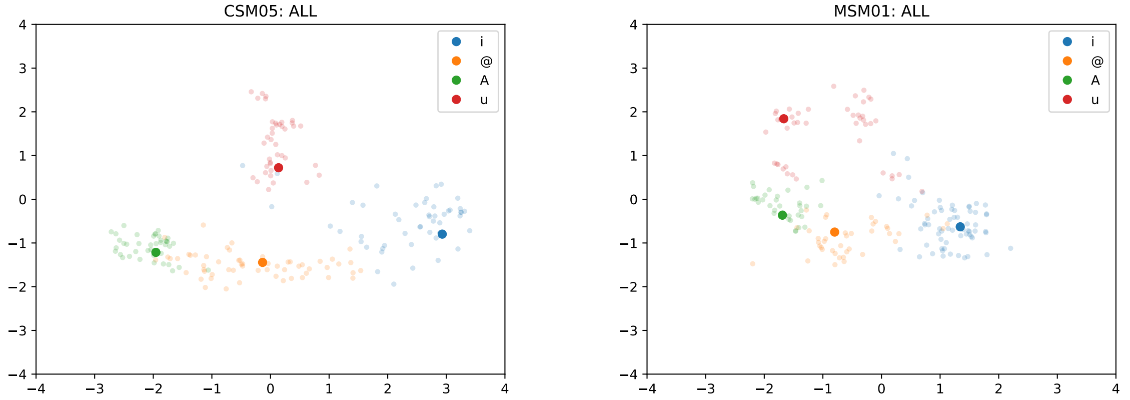
$$B = \frac{2}{C(C-1)} \sum_{i=1}^{C} \sum_{j=1}^{i-1} b_{i,j} \tag{2}$$

$$b_{i,j} = \frac{1}{N_i N_j} \sum_{k \in C_i} \sum_{l \in C_j} \|\mathbf{x}_k - \mathbf{x}_l\| \tag{3}$$

and W is the average pair-wise Euclidian distance of the data *within* each class

$$W = \frac{1}{C} \sum_{i=1}^{C} w_i \tag{4}$$

$$w_i = \frac{2}{N_i(N_i - 1)} \sum_{k \in C_i} \sum_{l \in C_i} \|\mathbf{x}_k - \mathbf{x}_l\|,\ k < l \tag{5}$$



(a) *most intelligible male speaker (87% transcription intelligibility)* (b) *least intelligible male speaker (29% transcription intelligibility)*

Figure 2: *Two-dimensional PCA of MFCC features for peripheral vowels. Note that classes are better separated for the more intelligible speaker. Larger dots represent the class centroids.*

with N_i representing the number of data points for class i. The proposed similarity measure is similar to Fisher's Discriminant Ratio, but makes no assumptions on the underlying data distribution. It is invariant under transformation, scaling, and rotation (in two- and three-dimensional space). Note that this measure is independent of the type of feature used.

We applied the proposed measure to all formant vectors associated with instances of a vowel; i. e. C=4. Then we predicted S→SI and S→TI, and found r^2=0.19 (p<0.01) for both.

5. Vowel Spectral Metric

In this section, we replace formant features with short-term spectral features. The advantage of such a representation is that it does not require manual correction; moreover, it can capture spectral details beyond peak locations. Specifically, we used 13th-order Mel-frequency cepstral coefficients (MFCC) to represent the spectrum; this type of feature has been successfully used in automatic speech recognition and speaker identification system, as it (1) removes spectral effects of the fundamental frequency, (2) applies a perceptually-relevant frequency scale, and (3) provides a reasonable approximation of the detailed spectrum using orthogonal components. We excluded the zeroth MFCC (reflecting the total energy of the spectrum), thus solely focusing on the spectral shape; however, we plan on using a normalized energy value in the future.

Fig. 2 shows a projection of the feature data onto two dimensions for the purpose of visualization, using principal component analysis (PCA), using the most and least intelligible male speaker for comparison. We calculated the separability S (see Section 4.2) of the MFCC features for the peripheral vowels for each speaker and then fit linear models predicting S→SI, and S→TI intelligibility. The corresponding r^2 values were 0.22 (p<0.01), and 0.29 (p<0.01), respectively.

6. Unsupervised Spectral Metrics

In this section, we will consider metrics that do not require the existence of prior phonetic labeling, and which utilize the entire speech signal. We continue using MFCC features to model the short-term spectral trajectory. We will examine two approaches.

6.1. Minimum Spanning Tree based entropy

Similar to recent work by Jiao et al. [10], we explored using entropy as a measure of variation. We first constructed a fully-connected Euclidean graph from $\mathbf{X}_{N\times D}$. A spanning tree $\tau_\mathbf{X}$ is a connected acyclic graph passing through all N points in $\mathbf{X}$, with edges $e_{i,j}$ connecting pairs $(\mathbf{x}_i, \mathbf{x}_j)$, and associated lengths, denoted as $|e_{i,j}| = \sqrt{(x_{i,1} - x_{j,1})^2 + \ldots + (x_{i,D} - x_{j,D})^2}$. The power-weighted length of the tree is the sum of all edge lengths raised to a power γ

$$L(\tau_\mathbf{X}) = \sum_{e \in \tau_\mathbf{X}} |e|^\gamma \tag{6}$$

We find a *minimal* spanning tree (MST) $\tau^*_\mathbf{X}$ that has the minimum length of all possible spanning trees. An estimate of the order υ Rényi entropy [16] can then be defined as

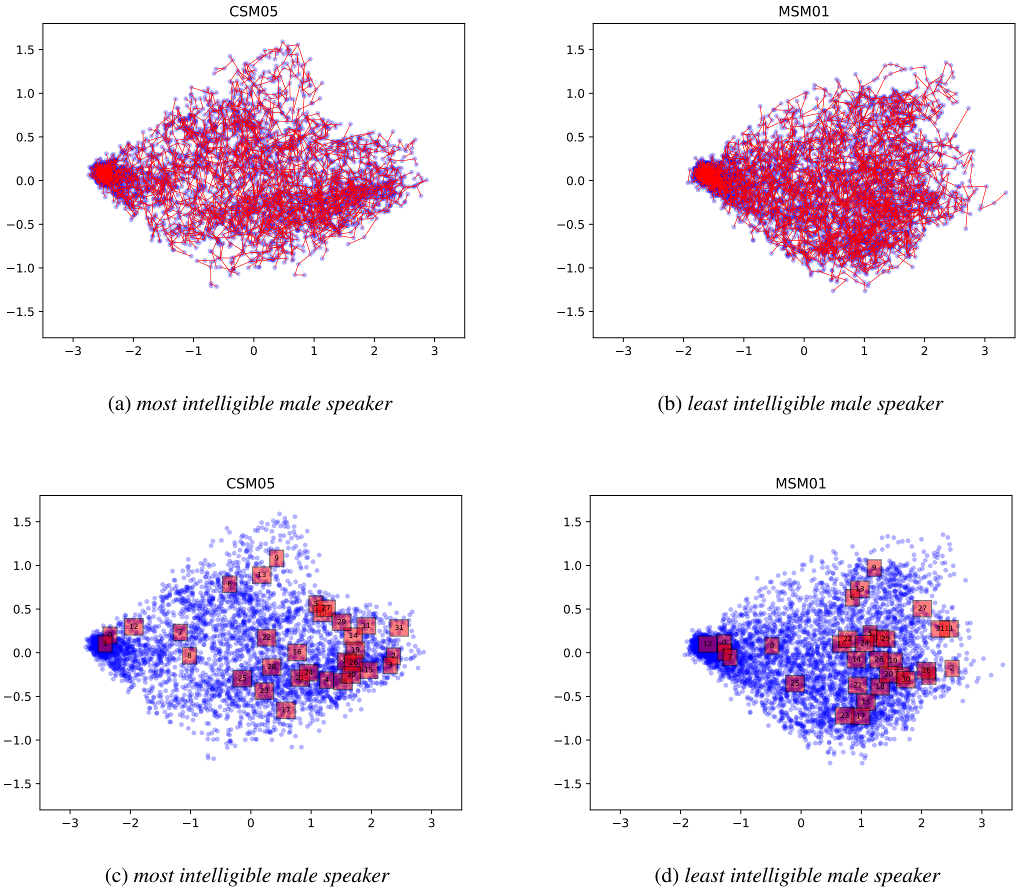
$$\hat{H}_\upsilon(\mathbf{X}) = \frac{1}{1-\upsilon} \ln N^{-\upsilon} L(\tau^*_\mathbf{X}) \tag{7}$$

where we use $\upsilon = 0.99$ (the limiting value of H_υ as $\upsilon \to 1$ is the Shannon entropy) and $\gamma = D \cdot (1 - \upsilon)$. Using bootstrapping with subsets of $\mathbf{X}$ (on the sentence-level, not the frame-level), we obtained relatively consistent and robust estimators of Eq. 7 [17, 18]. Figures 3a and 3b show two examples.

We applied the measure to the MFCC features and predicted H→SI and H→TI at r^2<0.01 (p>0.6) for both, surprisingly. In comparison, Jiao et al. [10] used H to predict the perception of articulatory precision of 57 speakers, resulting in r^2=0.34 (a p value was not given). However, the study differed form ours in several respects: (1) articulatory precision is not the same as SI or TI measures, (2) the study contained speakers with mild to severe dysarthria, (3) articulatory precision was estimated using only seven listeners, and (4) the type of feature that was used was different from the type of feature used here.

6.2. Hidden Markov Model state separability

We propose an alternative approach to measuring variation without prior class information via clustering the data to determine classes, and then using the class information to measure separability. For clustering, we propose to use a hidden Markov



(a) *most intelligible male speaker*

(b) *least intelligible male speaker*

(c) *most intelligible male speaker*

(d) *least intelligible male speaker*

Figure 3: *The top panel shows the minimum spanning trees after PCA. (Note that because of the projection the tree appears to be invalid.) The bottom panel shows the HMM means ($Q = 32$) after fitting.*

model (HMM). The key idea is that each state models a region in feature space that can loosely be associated with a context-free phoneme.

The initial speaker-independent HMM was constructed as follows: Using a separate database [19], the HMM was initialized from a trained general mixture model (GMM) with Gaussian distributions, which in turn were initialized from a k-means clustering algorithm. The prior transition matrix was ergodic, with an emphasis on self-loops, and starting and ending states favored the class representing pauses. The HMM was then trained in an unsupervised manner (see Figures 3c and 3d).

To perform clustering of one of the speakers in our corpus, we adapted the speaker-independent HMM by continuing training using the sentences from our corpus, separately for each speaker. After training finished, we decoded the sentences, effectively creating Q classes. We then applied our proposed separability measure (see Section 4.2). For $Q = \{20, 24, 28, 32\}$, for S→SI we obtained r^2 values between 0.13 and 0.18 (for Q=20, p<0.01), and for S→TI we obtained values between 0.10 and 0.13 (for Q=20, p=0.03).

7. Conclusion

We compared the degree to which a set of objective measures are capable of predicting speech intelligibility. We found that a VSA-based measure using automatic-formant frequency tracks has comparable performance to one with manually-corrected tracks. However, a proposed class-separability (where classes were vowels) measure using the same data resulted in higher r^2 values. Even better values were found when replacing formant features with short-term spectral features in the form of MFCCs. We then explored unsupervised approaches that do not require phonetic labeling. We found that a previously-proposed minimum spanning tree based entropy measure did not yield useful results when used with our data and feature type. In contrast, we obtained low r^2 values when using the class-separability measure, where classes were given by the segmentation of a HMM that was fit to the data without supervision. In planned comparisons, we found that only one comparison was statistically significant (p<0.05), namely the vowel spectral metric vs. the entropy measure, at $p = 0.03$. We plan on using additionally available speaking styles of the corpus to address this. Overall, it seems that a measure of separability, i. e. the ratio of between-class variance vs. within-class variance, may be a better predictor of speech intelligibility than global variance in the form of VSA or entropy. These results are comparable to other dysarthria studies that arrived at similar conclusions [3, 6]. In closing, development of fully-automated metrics predictive of intelligibility will facilitate adoption of objective acoustic tools in clinical practice.

8. References

[1] S. Skodda, W. Visser, and U. Schlegel. Vowel articulation in parkinson's disease. *Journal of Voice*, 2010.

[2] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox. Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *J. Speech Lang. Hear. Res.*, 53:114–125, 2010.

[3] Heejin Kim, Mark Hasegawa-Johnson, and Adrienne Perlman. Vowel contrast and speech intelligibility in dysarthria. *Folia Phiatr Logop*, 63:187–194, 2011.

[4] Fredrik Karlsson and Jan van Doorn. Vowel formant dispersion as a measure of articulation proficiency. *J. Acoust. Soc. Am.*, 132 (4):2633–2641, 2012.

[5] Kris Tjaden, Jennifer Lam, and Greg Wilding. Vowel acoustics in parkinson's disease and multiple sclerosis: Comparison of clear, loud, and slow speaking conditions. *J Speech Lang Hear Res*, 56: 1485–1502, 2013.

[6] Kaitlin L. Lansford and Julie M. Liss. Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research*, 57:57–67, February 2014.

[7] Annalise R. Fletcher, Megan J. McAuliffe, Kaitlin L. Lansford, and Julie M. Liss. Assessing vowel centralization in dysarthria: A comparison of methods. *Journal of Speech, Language, and Hearing Research*, 2017.

[8] Brad H. Story and Kate Bunton. Vowel space density as an indicator of speech performance. *J. Acoust. Soc. Am.*, 141, 2017.

[9] Steven Sandoval, Visar Berisha, Rene L. Utianski, Julie M. Liss, and Andreas Spanias. Automatic assessment of vowel space area. *J. Acoust. Soc. Am.*, 134(5), 2013.

[10] Yishan Jiao, Visar Berisha, Julie Liss, Sih-Chiao Hsu, Erika Levy, and Megan McAuliffe. Articulation entropy: An unsupervised measure of articulatory precision. *IEEE Signal Processing Letters*, 2017.

[11] K. Tjaden, J. E. Sussman, and G. E. Wilding. Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in parkinson's disease and multiple sclerosis. *J Speech Lang Hear Res*, 57(3):779–792, 2014.

[12] K. L. Stipancic, K. Tjaden, and G. Wilding. Comparison of intelligibility measures for adults with parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research*, 59(2):230–238, 2016.

[13] K. Yorkston, D. Beukelman, and R. Tice. *Sentence Intelligibility Test for Macintosh*. Communication Disorders Software, Lincoln, NE, 1996.

[14] Ieee recommended pratice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, September 1969.

[15] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[16] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pages 547–561, 1961.

[17] A. O. Hero and O. Michel. Robust entropy estimation strategies based on edge weighted random graphs. In *SPIE International Symposium on Optical Science, Engineering and Instrumentation*, San Diego, 1998.

[18] A. Hero and O. Michel. Asymptotic theory of greedy approximations to minimal k-point random graphs. *IEEE Trans. on Inform. Theory*, 1999.

[19] John Kominek, Alan W Black, and Ver Ver. Cmu arctic databases for speech synthesis. Technical report, CMU, 2003.