INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Unsupervised Methods for Evaluating Speech Representations

Michael Gump, Wei-Ning Hsu, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA

{gump, wnhsu, glass}@csail.mit.edu

Abstract

Disentanglement is a desired property in representation learning and a significant body of research has tried to show that it is a useful representational prior. Evaluating disentanglement is challenging, particularly for real world data like speech, where ground truth generative factors are typically not available. Previous work on disentangled representation learning in speech has used categorical supervision like phoneme or speaker identity in order to disentangle grouped feature spaces. However, this work differs from the typical dimension-wise view of disentanglement in other domains. This paper proposes to use low-level acoustic features to provide the structure required to evaluate dimension-wise disentanglement. By choosing well-studied acoustic features, grounded and descriptive evaluation is made possible for unsupervised representation learning. This work produces a toolkit for evaluating disentanglement in unsupervised representations of speech and evaluates its efficacy on previous research.

Index Terms: speech representation learning, unsupervised learning

1. Introduction

Disentanglement is a heuristic thought to be important in learning interpretable and informative representations [1]. Informally, disentanglement is the property that changes in individual axes of a representation correspond to changes in individual ground truth factors of the observed data. While this idea is intuitively appealing it has been difficult to evaluate in practice, and is particularly difficult for real world data. Since there may be infinitely many generative factors that could explain the data, it is not possible in general to learn models that are disentangled with the "right" generative factors. Therefore, metrics proposed to evaluate the level of disentanglement in a representation typically require the use of ground truth generating factors [2, 3, 4]. Conditioned on the generating factors, the observations should have very little variation.

In computer vision, simulated data is often used in order to make these ground truth generative factors available for analysis. The Chairs [5], 3D-Faces [6], and dSprites [7] datasets are often used in research concerning disentanglement. Each of these are procedurally generated from a set of factors, for instance the dSprites dataset is a set of 2D shapes generated from color, shape type, scale, rotation, and position. Then, a representation can be shown to be well-disentangled by demonstrating that traversing any one of these factors causes changes in only a single dimension of the learned latent space.

For real world data like speech, this paradigm is difficult to achieve since the ground truth generative factors are not known. Existing work on disentanglement in speech [8, 9, 10] has considered high-level factors like phoneme category or speaker identity as a source of structure. However, even when conditioned on both speaker and phoneme a significant amount of variation in segments is left unexplained. Since this makes it difficult to evaluate dimension-wise disentanglement, these methods have instead focused on the disentanglement between groups of latent dimensions. For example, the factorized hierarchical variational autoencoder (FHVAE) [10] disentangles latent factors representing utterance-level variation from latent factors representing the remaining segment-level variation. The disentanglement between these two feature spaces can then be evaluated using speaker and phoneme labels respectively.

Quantifying dimension-wise disentanglement for real world data is still an unsolved problem and would require more granular signal about the generative process. This paper proposes a methodology for evaluating dimension-wise disentanglement for real world data by grounding evaluation with techniques in acoustic processing. This approach has two potential benefits over previous work. First, traditional acoustic processing has the potential to describe most of the variation in speech using interpretable features, enabling more grounded evaluation. In addition, the approach of this paper is completely unsupervised meaning that it can be applied on a variety of large-scale datasets, though the acoustic feature extraction may need to be tuned to particular speech domains.

This paper evaluates the effectiveness of the described evaluation techniques on FHVAE, which has previously been shown to learn dimension-wise disentangled representations [11]. For this model we demonstrate how quantitative and qualitative methods can be used to assess the degree of dimension-wise disentanglement. In addition, we show that using acoustic features can allow interpretable insights to be made about speech representations. Evaluation is preformed using TIMIT [12], a read speech corpus designed to provide a rich set of phonetic contexts, with 630 speakers and 5 hours of data. Lastly, we release[1] an evaluation suite in the hope of facilitating future research in disentangled representation learning for speech.

2. Grounding Evaluation with Acoustic Features

One set of ideal factors for evaluating dimension-wise disentanglement for speech would be articulatory parameters like vocal tract length, airflow strength, or tongue position, since these parameters would explain most of the variance of the data. However, this ground truth data is difficult to measure [13, 14] and even more challenging to estimate reliably [15, 16]. Thus, these factors are out of reach for analysis on large-scale datasets. Fortunately, previous research has shown that variation in these factors can be explained by the formant peaks and fundamen-

[1] https://github.com/mhgump/acoustic_disentanglement

http://dx.doi.org/10.21437/Interspeech.2020-2990

Feature	σ	σ_{phone}	σ_{vowel}
F1	818.7	549.7	256.9
F2	1129.2	805.3	695.2
F3	1291.7	1116.9	947.1
F1 bw	201.1	153.6	100.8
F2 bw	232.8	215.3	185.6
F3 bw	255.4	249.3	241.2
F0	102.7	549.7	256.9
Energy (E_{hatn})	0.142	0.069	0.200
Magnitude ($M_{\hat{n}}$)	2.241	1.248	2.668
Zero Crossing Rate	0.169	0.079	0.037

Table 1: *The standard deviation of several features extracted from TIMIT. The variance is computed over each phoneme and the square root of the average is reported as σ_{phone}. σ_{vowel} is analogous but only considers vowels.*

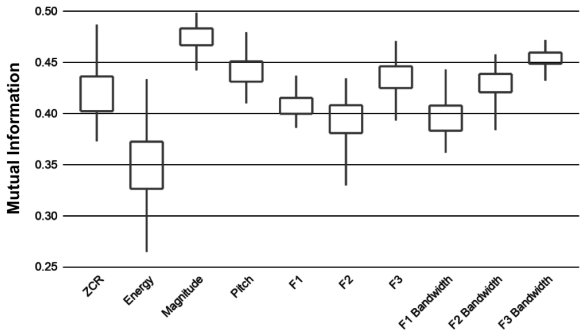

Figure 1: *Box plots showing range of mutual information between mel-spectrogram and each acoustic factor, across the hyper-parameters for discretization.*

tal frequency, among other significant acoustic features. Traditional rules based synthesizers and text-to-speech systems use these features to reconstruct speech [17, 18, 19, 20]. Additionally, there are reliable and well-studied algorithms for estimating these features [21, 22]. Lastly, the connection between the extracted factors with distinctive features like voicing or place/manner features have been the subject of considerable research [23], so they provide a very interpretable structure for speech scientists.

This paper considers fundamental frequency, energy, magnitude, and formant peaks and bandwidths as perceptually significant features for analysis. We consider three different methods for estimating fundamental frequency, RAPT [24] and SWIPE [25] are knowledge based pitch tracking algorithms, and CREPE [26] is a data driven method. We do not use a single estimation method since each method has different failure modes, our final raw pitch value is the average of the individual estimates. We estimate the first three formant peaks and bandwidths using LPC coefficients [27]. Formant trackers that incorporate frame to frame information are not considered since they are sensitive to gender and are typically tuned to one or the other [28]. Higher formants were not included because they are often not very distinct and therefore hard to estimate reliably. Table 1 shows the standard deviation of each feature across TIMIT in order to loosely describe the reliability of these features. For TIMIT, phonetic transcriptions are available so we additionally report the consistency of each feature when grouped by phoneme label and when only frames of vocalic sounds are considered. This demonstrates a pattern we expect to see, that F1 and F2 are very consistent for the same vowel.

Generative models in speech typically are trained to encode multiple frames of short time input features such as MFCCs or the mel-spectrogram. By learning to represent multiple frames the models are able to learn sub-phone or phone-like units which would not be possible from individual frames. Because of this it is not sufficient for representations to be compared to the average factor value extracted from the segment of speech it encodes. Since the representation has actually learned to encode a *trajectory* of values. In this paper, we represent the trajectories of each segment more granularly by extracting both the factor mean and factor standard deviation from each segment. The extracted factor values are continuous but as a post-processing step we discretize them. Each factor is binned so that each bin corresponds to a range of segment means and a range of segment standard deviations. Discretization simplifies the computation of metrics like mutual information and EER. Additionally, it smooths noise from the estimation process making visualizations more effective. We would like to verify that the discretization process produces clusterings of the data that are relatively insensitive to the choice of hyper-parameters. Discretization for a factor depends on the number of ranges chosen for the segment means and the number of ranges chosen for the segment standard deviations, the product of these numbers is the total number of bins. The statistics of each feature value are used to compute appropriate bin edges. In Figure 1, the mutual information between the a 80 dimensional mel-spectrogram representation is computed as the hyper-parameters are varied between pairs that would produce between 25 and 2500 total bins. The boxplots demonstrate that the difference across factors is more significant then the difference of hyper-parameters in this range. When applied to new datasets, the number of total bins should be chosen to ensure that the number of samples in each bin is relatively constant as otherwise metric values can be increased without bound.

3. Evaluation Suite for Unsupervised Representations

In order to enable future research into dimension-wise disentanglement of speech representations, an evaluation suite is implemented and released in Python. The toolkit includes a variety of qualitative and quantitative analyses alongside acoustic processing pipelines. Users are expected to train and extract speech representations before testing, the tool can then be used to produce visualizations and metrics for each representation. The representations do not need to be computed at the same window length and stride length as our acoustic features, as the tool will interpolate to compute statistics over the appropriate segments. If the representations correspond to uneven or randomly sampled segments of utterances, then the segment positions can be provided to the tool.

The suite natively supports the TIMIT corpus and has the required data processing pipelines for parsing the standard format. Running analysis over other datasets is simple and merely requires that the user specify the directory structure for locating waveform files. Transcriptions or other types of data can also be made available to the toolkit but requires the necessary parsing to be implemented by the user, examples are given to

Feature	σ	σ_{phone}	σ_{vowel}
F1	588.4	553.6	544.6
F2	625.7	607.4	604.1
F3	837.5	731.7	708.0
F0	155.7	155.9	158.1

Table 2: *The standard deviation of several features extracted after reconstruction from TIMIT. The variance is computed over each phoneme and the square root of the average is reported as σ_{phone}. σ_{vowel} is analogous but only considers vowels.*

assist in this process. As motivated by the previous section the toolkit contains acoustic processing for several features such as formant, pitch, and energy. Additional features can be added by the user to take advantage of the later stages of the toolkit's pipeline.

3.1. Quantitative Methods

Three quantiative metrics are implemented to make use of acoustic feature extraction. [4] describes a method for estimating the mutual information (MI) between a parameterized posterior latent distribution $p(z|x)$, which is usually Gaussian, and a set of discrete ground truth latent factors v. From this estimate of MI they also propose the mutual information gap (MIG) metric, which measures disentanglement. MIG computes the expected gap in MI between the highest scoring latent dimension and the second highest, intuitively this penalizes representations that encode generative factors in more than one dimension. The equal error rate (EER) for predicting the correct discrete generative factor is computed from the representation's similarity matrix using cosine distance.

3.2. Analysis of Reconstructed Data

So far the use of acoustic processing has been focused only on the relationship between a representation and the data that it encodes. For deep generative models, this only describes the inference side and we may be interested in analyzing the generative model. In particular, previous work has attempted to verify that changes in a latent dimension correspond to specific changes in the output of the generative model [11]. This requires that we can estimate the acoustic features from the spectrum features generated by the model. The Griffin-Lim audio reconstruction method is used to allow this type of analysis. Since it is difficult to reconstruct audio, particularly from synthesized spectral features we should expect this method to be noisy. Table 2 reproduces Table 1 using audio reconstructed from 20 dimensional MFCCs in order to demonstrate the feasibility of this method. By grouping the data by phoneme we should expect the variance in formant peaks to decrease, particularly for vowels where the formants are predictable. Since this no longer occurs for reconstructed features we can see that there has been an increase in estimation error from Griffin-Lim.

We describe a method for estimating the relationship between a latent variable and acoustic features estimated from the output of a generative model. For a given dimension k of the representation and M probe points p_i according to the prior $p(z)$, we sample $z \sim \mathbf{E}_{x \sim X}[p(z|x)]$ and generate M latent vectors where all dimensions are fixed except z_k which is varied along the probe points. All latent vectors can then be fed to the generative model and the sampled spectral features can be used to reconstruct audio. This can be used to construct a plot of the average feature value or as input to some quantitative metric like mutual information.

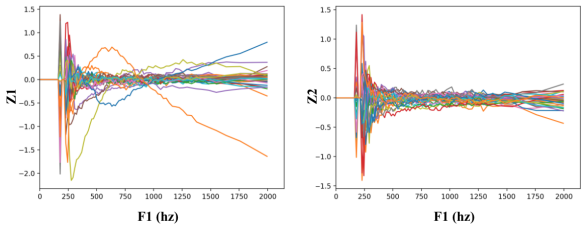

Figure 2: *Both z_1 and z_2 are compared to F1, several dimensions of z_1 take extreme values due to change in F1 and no such trend is seen for z_2. Spikes at low frequencies are due to small numbers of samples.*

4. Experiments

This section describes the methods used to test the evaluation suite. The factorized hierarchical variational autoencoder (FHVAE) [29, 30] is considered because it learns to encode segment-level and utterance-level factors into different latent variables (z_1 and z_2 respectively). In addition we consider the means of these latent variables aggregated across each utterance as μ_1 and μ_2. The evaluation suite allows us to demonstrate several interesting properties about these representations. The FHVAE model used has two LSTM layers of 256 cells each for all encoder and decoder networks, we used a discriminative weight of $\alpha = 10$ and learn latent spaces each with 32 units each. The model was trained on TIMIT [12], with 240 utterances from 24 speakers reserved for evaluation.

4.1. Visualizations

In order to verify if either of z_1 or z_2 are dimension-wise disentangled with respect to any of the extracted acoustic features, we plot trends between the factor value and the magnitude of each dimension of the latent spaces. In Figure (2), we do this for the first formant peak (F1). For each plot, the dataset is partitioned using the segment-level means of F1. Then, each dimension of the latent space is plotted as a line representing the average magnitude of the dimension within the corresponding partition of the dataset. For the plot corresponding to z_1, we see that only a few dimensions have a relationship with F1, indicating dimension-wise disentanglement. For z_2, no such relationship can be seen with F1. The noise at low frequencies for each plot is due a small number of samples in those bins. The results in these plots confirm what we would expect. F1 typically varies segment to segment and z_1 encodes segment-level variation while z_2 is constrained to ignore such variation.

To further investigate the relationship between the z_1 latent space and segment-level variation, we attempt to determine if it is sensitive to phonetic categories or other linguistically significant cues. To do so we compare the intensity of each dimension of z_1 with the segment means and standard deviations for both F1 and F2. Figure 3 shows the results of this for two dimensions of z_1. In the top left, the dataset is partitioned by both the segment mean and segment standard deviation of F1 and the pixels display the average intensity of the 15^{th} dimension for samples from the respective partition. This process is repeated in each

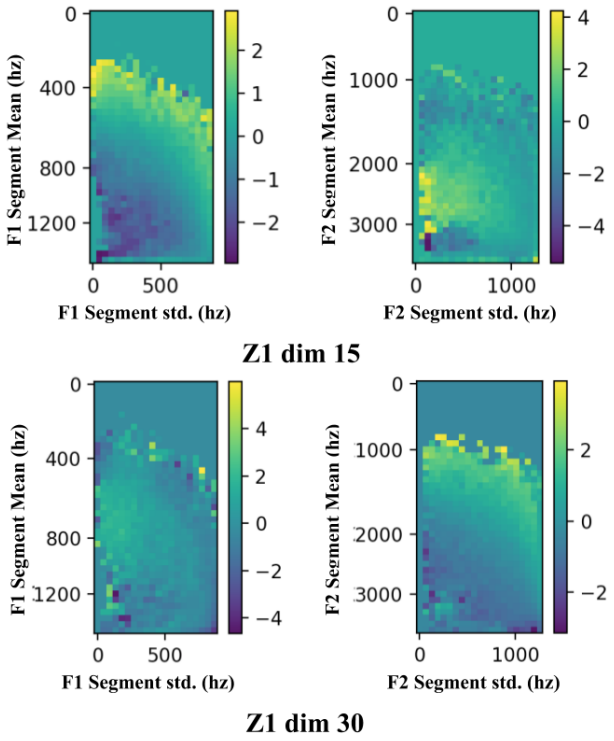

Figure 3: *Two dimensions of z_1 are plotted with respect to F1 and F2. Each bin corresponds to a range of segment means and segment standard deviations.*

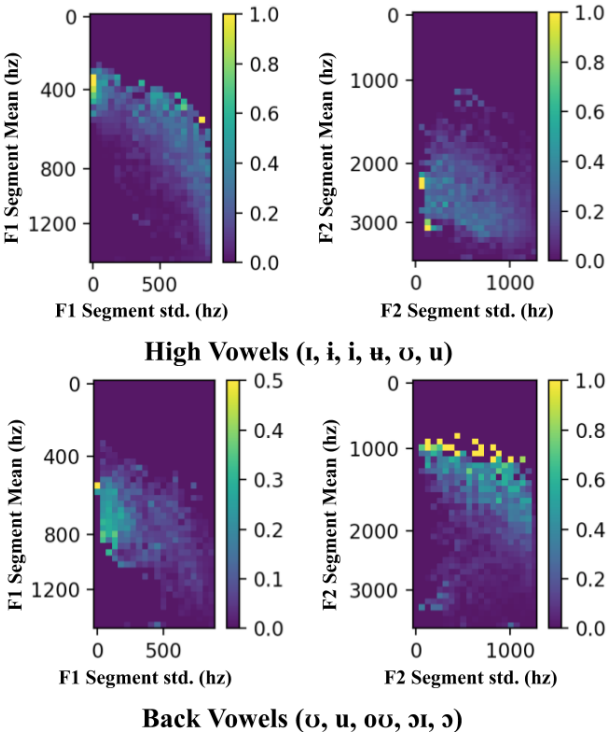

Figure 4: *The frequency of two phonetic categories are plotted with respect to F1 and F2. Each bin corresponds to a range of segment means and segment standard deviations. Note the correspondence with Figure 3.*

quadrant for the appropriate feature, either F1 or F2, and the appropriate dimension. We can see that the 15^{th} dimension is sensitive to low F1 values and high F2 values, while the 30^{th} seems sensitive to low F2 values. This process is useful for quickly characterizing dimensions of a representation in terms of grounded features. After manual analysis of the visualizations produced for z_1 for all acoustic features, other dimensions were found to be sensitive to a variety of conditions, such as voicing as determined by pitch.

Using our linguistic knowledge, we may suspect that these dimensions are not just sensitive to arbitrary patterns in the data, but real phonetic categories. Indeed, high vowels, with low F1 and distributed F2, would seem to correspond to dimension 15 and back vowels (low F2, low to mid F1) with dimension 30. In Figure 4 we produce analogous plots that measure the frequency of the corresponding phonetic categories instead of latent intensity. We find that the produced plots match the latent intensities very closely, indicating that our hypothesis was correct. Note that the scales of these axes are different since these are probabilities scaled from 0 to 1 rather then real valued intensities. It iss important to note that this identification of phoneme sensitive latent dimensions would not be possible if only the segment means were considered. Since an increase in formant mean for a segment may correspond with an increase in the variance as well, trends over a single dimension do not display the patterns as well. More detailed statistics for each feature trajectory have the potential to be even more discriminative.

5. Conclusion

This paper discusses methods for treating acoustic features as ground truth generative factors for the purpose of evaluating unsupervised representation learning. This methodology provides the necessary structure for evaluating dimension-wise disentanglement in speech, both quantitatively and qualitatively. In addition, the evaluation methods allow grounded and interpretable analysis for speech representations. We validate previous understanding of FHVAE using these methods and perform novel analysis of the representations it produces. The tools described in this paper are made available as an open-source toolkit written in Python. The hope is to enable comparison of methods in disentangled representation learning in the complex and real-world domain of speech. Future work should focus on validating the reliability of quantitative metrics computed using the extracted acoustic factors.

6. References

[1] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR*, vol. abs/1206.5538, 2012. [Online]. Available: http://arxiv.org/abs/1206.5538

[2] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[3] H. Kim and A. Mnih, "Disentangling by factorising," 2018.

[4] T. Q. Chen, X. Li, R. B. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *CoRR*, vol. abs/1802.04942, 2018. [Online]. Available: http://arxiv.org/abs/1802.04942

[5] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic, "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *CVPR*, 2014.

[6] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, ser. AVSS '09. USA: IEEE Computer Society, 2009, p. 296–301. [Online]. Available: https://doi.org/10.1109/AVSS.2009.58

[7] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dsprites: Disentanglement testing sprites dataset," https://github.com/deepmind/dsprites-dataset/, 2017.

[8] S. Khurana, S. Joty, A. Ali, and J. Glass, "A factorial deep markov model for unsupervised disentangled representation learning from speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6540–6544.

[9] Y. Li and S. Mandt, "Disentangled sequential autoencoder," 2018.

[10] W. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," 2017.

[11] W. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," *CoRR*, vol. abs/1810.07217, 2018. [Online]. Available: http://arxiv.org/abs/1810.07217

[12] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[13] A. Joshi and C. Watts, "A comparison of indirect and direct methods for estimating transglottal airflow rate," *Journal of Voice*, 11 2017.

[14] F. Chen, S. Li, Y. Zhang, and J. Wang, "Detection of the vibration signal from human vocal folds using a 94-ghz millimeter-wave radar," *Sensors*, vol. 17, no. 3, p. 543, Aug 2017.

[15] A. Patil and M. S. Shah, "Comparison of vocal tract shape estimation techniques based on formant frequencies, autocorrelation, covariance and lattice," in *2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE)*, 2015, pp. 1–6.

[16] A. Lammert and S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLOS ONE*, vol. 10, no. 7, pp. 1–23, 07 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0132193

[17] X. Rodet, "Time-domain formant-wave-function synthesis," *Computer Music Journal*, vol. 8, no. 3, pp. 9–14, 1984. [Online]. Available: http://www.jstor.org/stable/3679809

[18] R. Carlson and B. Granström, "Rule-based speech synthesis," *Springer Handbook of Speech Processing*, p. 429–436, 2008.

[19] N. Aoki, "Development of a rule-based speech synthesis system for the japanese language using a melp vocoder," in *2000 10th European Signal Processing Conference*, 2000, pp. 1–4.

[20] J. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, 1973.

[21] P. Harrison, "Making accurate formant measurements: an empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements," Ph.D. dissertation, University of York, 2013.

[22] D. Jouvet and Y. Laprie, "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1614–1618.

[23] T. Nearey, "Static, dynamic, and relational properties in vowel perception," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2088–2113, 1989. [Online]. Available: https://doi.org/10.1121/1.397861

[24] D. Talkin, "A robust algorithm for pitch tracking (rapt)," 2005.

[25] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, pp. 1638–52, 10 2008.

[26] J. Kim, J. Salamon, P. Li, and J. Bello, "Crepe: A convolutional representation for pitch estimation," 2018.

[27] D. Gargouri, A. Kammoun, and A. Hamida, "A comparative study of formant frequencies estimation techniques," 01 2006.

[28] F. Schiel and T. Zitzelsberger, "Evaluation of automatic formant trackers," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1449

[29] W. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *CoRR*, vol. abs/1709.07902, 2017. [Online]. Available: http://arxiv.org/abs/1709.07902

[30] W. Hsu and J. Glass, "Scalable factorized hierarchical variational autoencoder training," 2018.