Unsupervised Methods for Evaluating Speech Representations

by

Michael H. Gump

S.B., Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Electrical Engineering and Computer Science
May 18, 2020

Certified by .....
James R. Glass
Senior Research Scientist
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Unsupervised Methods for Evaluating Speech Representations

by

Michael H. Gump

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Unsupervised representation learning using deep generative models has produced remarkable results across many domains in recent years. These methods have been applied to speech processing to learn representations useful for downstream supervised tasks like speaker, dialect, or phoneme identification. One research path has been to develop general purpose priors that select effective representations. However, many priors on good representations are difficult to incorporate into unsupervised methods because they are difficult to evaluate without supervision. This thesis proposes to use low-level acoustic features to address this problem for speech. By using techniques in acoustic processing, we develop methods for structured evaluation for speech representations. The evaluation aims both to assess the efficacy of representations for downstream tasks and to validate claims about the priors used to construct them. An evaluation suite for benchmarking and analyzing research in speech representation learning is produced and open-sourced as a result of this thesis.

Thesis Supervisor: James R. Glass
Title: Senior Research Scientist

Acknowledgments

I'm extremely grateful to Professor Jim Glass and Wei-Ning for their generosity, valuable insights, and patience. Thank you for providing me with the resources necessary to pursue research I was interested in. I'd also like to thank the SLS group as a whole, where I've had many thoughtful and interesting conversations. Lastly, thank you to my family and friends who have been there for me throughout this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Challenges in Representation Learning

Unsupervised learning is an appealing paradigm because it promises to gain structured insight from whatever data is readily accessible. While traditional machine learning systems rely on large amounts of labeled, in-domain data, one aim of research in unsupervised learning is to make better use of more abundant unlabeled data. One approach in unsupervised learning is to find an effective way to represent the structure and variation of a data domain using unlabeled data, in other words representation learning. Representation learning is an increasingly rich line of research that attempts to find general-purpose heuristics for learning useful representations. Goals of the field have included learning more compact representations to reduce dimensionality, learning more interpretable representations to discover structure in a domain, and learning representations that are invariant to nuisance factors to improve performance on downstream tasks. To achieve these goals many heuristics have been hypothesized to be important, sparsity (Makhzani and Frey, 2013) and smoothness (Cemgil et al., 2020; Cai et al., 2019) are a couple of notable examples. (Bengio et al., 2012) discusses the field and various approaches. However, translating these priors into learning algorithms is difficult.

Priors on representations are often too loosely defined to guarantee that they will always select the representations that we are interested in. To remove this ambiguity,

either the heuristics in question must be more formally defined or useful biases on the data or models must be proposed. One example of this can be seen from research into disentanglement (Higgins et al., 2017; Kim and Mnih, 2018; Chen et al., 2018; Li and Mandt, 2018). A recent review (Locatello et al., 2018) of unsupervised studies on disentanglement provided several sobering results. Since there are infinitely many generative factors that could explain any dataset it is not possible in general to learn models that are disentangled with the "right" generative factors, and they proved a theoretical result regarding this. Empirically, they found significant disagreement between previously proposed metrics to measure disentanglement, suggesting that there is not yet consensus for how to reliably do so. Lastly, they found that model selection over the hyper-parameters and random seeds of existing methods seemed to be impossible without supervision. The conclusion of this work is that future research should show explicitly how biases on the data or model allow models with properly disentangled posteriors to be selected. Future research should show that learned representations are disentangled with respect to the true generative factors, and it should show results on a variety of real-world domains such as speech. However, the generative factors for real-world data may not be available for analysis making it difficult to evaluate disentanglement. This problem is not specific to disentanglement but represents a broader trend in representation learning. Universal priors are difficult to incorporate into unsupervised methods because they are difficult to evaluate without supervision.

1.2 Disentanglement

Disentanglement is a heuristic thought to be important in learning interpretable and informative representations. 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 some biases on the data or model are necessary. Therefore, metrics proposed

to evaluate the level of disentanglement in a representation typically require the use of ground truth generating factors (Higgins et al., 2017; Kim and Mnih, 2018; Chen et al., 2018). Conditioned on the generating factors, the observations should have very little variation.

In computer vision, simulated data is often used to make these ground truth generative factors available for analysis. The Chairs (Aubry et al., 2014), 3D-Faces (Paysan et al., 2009), and dSprites (Matthey et al., 2017) datasets are often used in research concerning disentanglement. Each of these is 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 (Khurana et al., 2019; Li and Mandt, 2018; Hsu et al., 2017a) 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) (Hsu et al., 2017a) 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 a more granular signal about the generative process. This thesis 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 inter-

pretable features, enabling more grounded evaluation. In addition, the approach of this thesis 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.

1.3 Approach

Acoustic features such as formant, pitch, short-time energy, etc. have well-studied impact on human perception of speech. The interrelationships between these factors and their connection with distinctive features like voicing or place/manner features have been the subject of considerable research (Nearey, 1989), so they provide a very interpretable structure for speech scientists. This thesis proposes that leveraging domain knowledge to provide additional structure to the data may be a promising approach, and we explore that idea for speech. We describe an evaluation suite for speech representations that applies standard acoustic processing techniques to provide interpretable and grounded insight. Our approach selects features known to be connected to human speech perception like formant and pitch and uses them to construct quantitative and qualitative analysis tools. The assumption is that good representations for speech are those that are highly related with these features. The evaluation tools will be made open-source on Github in order to allow for benchmarking and comparison of methods in speech representation learning.

The tool is designed primarily with deep generative models in mind. Methods in this space have applied neural networks as powerful function approximators to produce novel and intelligible speech (van den Oord et al., 2016). Additionally, frameworks that simultaneously infer a latent distribution, like the variational autoencoder (Kingma and Welling, 2013), benefit from developments in representation learning and could be better validated by a tool for structured evaluation. However, many of the methods in this thesis are sufficiently general to be applied to any representation learning technique and so the contributions are not specific to deep generative models.

We evaluate the effectiveness of the described evaluation techniques on FHVAE,

which has previously been shown to learn dimension-wise disentangled representations (Hsu et al., 2018). 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 performed using TIMIT (Garofolo et al., 1992), 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.

1.4 Thesis Contributions

This thesis explores the efficacy of using low-level acoustic features to aid in the evaluation of unsupervised representation learning methods for the speech domain. An open-source evaluation suite is produced to make the necessary tools and baselines from this approach available to future work in representation learning.

Chapter 2 presents background on acoustic phonetics and various methods for estimating important acoustic features. Chapter 3 briefly describes the datasets and tools used in this thesis.

Chapter 4 presents specific methods for using acoustic features to evaluate representations of speech and measures the reliability of these methods on two standard speech datasets. Chapter 5 applies these methods to FHVAE, producing baseline results, and describing existing work in speech representation learning through a new lens.

Chapter 6 extends this analysis to a general set of tools for evaluating representations of speech and introduces the open-source evaluation suite.

Chapter 7 concludes the thesis by summarizing its contributions and outlining possibilities for future work.

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

Chapter 2

Background: Acoustic Processing

Acoustic signal processing studies the physical properties of sounds and how signal processing techniques can be applied to describe the acoustic signal. Speech processing focuses on the features of acoustic signals that are relevant to either the articulatory process of speech or the perceptual process.

The spectrogram, a visual representation of the frequency space of a speech sound is a standard initial processing technique. Many speech sounds, particularly those that are voiced are periodic and lend themselves well to analysis in the frequency space. Other sounds, like stops (/p/ or /b/) or fricatives (/th/ or /f/), are not periodic and show up as vertical bands on the spectrogram since the energy is spread throughout the frequency space. The Mel-scale (Stevens et al., 1937) is a method proposed to scale the raw spectrogram to align with human perception of sound. A raw spectrogram will measure the energy in different frequency bands spaced evenly on the frequency axis. The Mel-scale adjusts this so that each band corresponds to a level that is perceptually equidistant from the adjacent levels. The Mel-spectrogram then refers to the Mel-scaled spectrogram. The Mel-frequency cepstral coefficients (MFCCs) are a derivative representation that applies further processing to the Mel-spectrogram. Two main advantages of the MFCC representations are that it decorrelates each dimension of the representation and can remove the most noisy components of the representation. The performance of machine learning models that use MFCCs as input features is typically slightly higher than those that use the Mel-spectrogram. However, many

methods prefer to use the Mel-spectrogram to avoid imposing unnecessary constraints on how the spectral signal should be represented.

One general problem in acoustic speech processing is that any feature estimated in windows over an utterance will have some estimation uncertainty due to the choice of window size. This trade-off comes from the fact that while we want to use short windows that are sensitive to changes in the speech signal but too short windows mean that there is not enough data to estimate the feature for that frame accurately. The difference between narrowband and wideband spectrograms illustrates this difference. Narrowband spectrograms have longer window sizes and therefore worse resolution across time, but better resolution in the frequency domain. Wideband spectrograms are exactly the opposite. Choosing a window size for acoustic analysis is a choice between these two types of resolution.

Energy, magnitude, and zero crossing rate have been used to differentiate voiced from unvoiced sounds (Jalil et al., 2013). Zero crossing rate is often considered a crude signal for the fundamental frequency. These are relatively simple computations with a single definition, so the main source of error is choosing an appropriate window size. Formant and pitch are two directly interpretable features that have long been used in speech processing systems, but they are more challenging to estimate reliably. Formants are the peaks of the frequency spectrum corresponding to the resonances and therefore shape of the vocal tract. They are characterized by the center frequency of the peak and the width of the peak referred to as the formant bandwidth. Formant peaks and bandwidths have been estimated using linear predictive coding (Snell and Milinazzo, 1993) and by cepstrum analysis (Fattah et al., 2008; Gagouri and Kammounand Ahmed Hamida, 2006). The main challenge for estimating formants is in discounting spurious peaks found from analysis of the frequency spectrum. Higher formants (F4, F5...) are particularly difficult to estimate because the higher frequency peaks are often not very distinct. Deep learning has also been applied to formant estimation (Dissen et al., 2019) but suffers from the disadvantage that in-domain data must be available. Formant trackers are tools that incorporate information from multiple frames in order to improve the estimates on individual frames. Formant trackers

often require estimates of neutral voice formant averages in order to represent the central tendency of a trajectory of formants. While these methods are successful (Schiel and Zitzelsberger, 2018) they introduce hyper-parameters that must be tuned to the type of speech data. Higher formants were not included because they are often not very distinct and are less directly related to speech perception (Harrison, 2013).

Pitch refers to the perception of tone in speech, which can be understood by asking listeners if a pure sinusoid with a given period has the same "pitch" as a given speech recording. Pitch estimation usually refers to estimating the fundamental frequency (F0) which is a physical property of sound that is known to be related to pitch. Pitch estimation is difficult because periodicity in the signal may not be due to voiced speech but from the vocal tract filter or background noise. Dealing with a variety of articulatory conditions and tracking changes in F0 reliably is not a solved problem. RAPT (Talkin, 2005) is an influential method that generates many sets of candidate pitch tracks and uses dynamic programming to find the trajectory that best matches our priors about variation in F0. YIN (de Cheveigné and Kawahara, 2002) is a method based on autocorrelation. SWIPE (Camacho and Harris, 2008) is a modification of autocorrelation based methods that addresses several problems with them. CREPE (Kim et al., 2018) is a data driven method that uses machine learning to produce accurate pitch estimates. It was shown to be more accurate than several knowledge based pitch trackers but was evaluated in relatively controlled conditions. As with DeepFormant, a downside of data driven methods is that performance may not translate well across domains. (Jouvet and Laprie, 2017) reviews the performance for many of these methods and others in a variety of speech conditions.

Typical machine learning methods in speech processing extract representations like the Mel-spectrogram of MFCCs by striding a relatively short-time window ($\sim$20ms) over an utterance. This short-time window is not long enough for representations to capture detailed linguistic context, instead, they describe the very local properties of the recorded sound wave. So these short-time representations are usually concatenated together as sampled "segments" from the dataset. Some methods (Reynolds et al., 2000) apply additional processing to ensure that the extracted segments are

not silent. But this type of constraint is not always applied. Thus, the segments extracted for training a speech model often do not align directly with the phonetic units present in each utterance. As a result when clustering is applied to the learned representations, sub-phoneme or phoneme-like units as learned as opposed to units that correspond exactly to phoneme boundaries.

Chapter 3

Data & Tools

3.1 TIMIT

TIMIT (Garofolo et al., 1992) is a dataset of read English speech recorded at 16000 Hz designed to provide a rich set of phonetic contexts for the development of ASR systems. The corpus contains speech from 630 speakers, each reading 10 sentences selected from 2342 unique sentence prompts. In total, the dataset contains around 5 hours of audio and includes time aligned phonetic and word-level transcriptions. All of the speakers are labeled by gender and by one of 8 dialectal classifications based on the geographic region of the U.S. where they spent their childhood.

The sentences selected for each speaker were chosen to increase dialectal and phonetic variety in the dataset. Two sentences were designed to demonstrate the dialectal differences between the recorded dialects and were read by all speakers. Another set of sentences was designed to cover as many valid phoneme combinations and phonetic contexts as possible. Each of these sentences was read by several speakers. Lastly, each speaker read three unique sentences selected from text sources to increase the diversity of allophones produced by each speaker.

3.2 Librispeech

Librispeech (Panayotov et al., 2015) is a collection of audiobook recordings that is around a 1000 hours of 16000 Hz speech. This data is pulled from a much larger corpus of unaligned audiobook recordings. A two stage alignment process that filtered out recordings that deviated significantly from the transcripts produced the final approximately 1000 hours of data.

A model trained on the standard WSJ dataset (Garofolo et al., 2016) was used to separate the dataset into two noise conditions. Speakers were assigned to either the larger noisy partition or to the "clean" partition based on the word error rate (WER) of the WSJ model on their speech. Two "clean" datasets are available, one with 100 hours of data ("train-clean-100") and the other with 360 hours of data ("train-clean-360"). The "noisy" dataset ("train-other-500") is about 500 hours of data. Test and development sets of around five hours are available for both the noisy and clean conditions.

3.3 Tools

The acoustic processing described in this thesis relies on two open-source python packages for audio analysis: LibROSA and PySPTK. These packages provide implementations for speech processing fundamentals like the Mel-spectrogram or linear predictive coding (LPC).

3.4 Factorized Hierarchical Variational Autoencoder

The factorized hierarchical variational autoencoder (FHVAE) (Hsu et al., 2017a) has been shown in previous research to encode both segment-level and utterance-level information in separate latent spaces. Since the properties of the representation have been studied in several previous works (Hsu et al., 2018; Shon et al., 2018; Hsu and Glass, 2018), it is a good candidate for testing the evaluation methods of this thesis. FHVAE learns two latent spaces following a hierarchical graphical model, z_1 and

z_2. z_2 is constrained to be consistent within an utterance but discriminative across different utterances. z_1 encodes the remainder of the segment-level variation necessary to model the speech signal. The model also produces sequence-vectors, μ_1 and μ_2, the averages of the respective latent distributions. These representations lose the ability to capture segment-level variation but are a more predictive signal for utterance-level factors, particularly in the case of μ_2 which was shown to outperform the famous i-vector method (Dehak et al., 2011) as a representation for speaker verification.

Chapter 4

Methods

The goal of representation learning is to determine feature spaces that effectively capture the variation of a given domain. The most basic interpretation of this is to learn features that encode the data accurately, however accurate reconstruction is usually not our only goal. We hope that by choosing clever biases for our models and objectives we can learn representations that encode the "important" variation. For example, a representation might be tailored to a particular supervised task by picking out only the informative features and ignoring nuisance features. This thesis is concerned with evaluating which variation a representation encodes and how. To do so, representations must be compared to reliable signals about the speech that they encode. In this chapter, we discuss how acoustic processing can be used to extract a useful signal for evaluating speech representations. Then we describe how to use this signal to evaluate the content of a representation and to evaluate disentanglement.

We consider 10 different features in order to approximate the true generative process of speech: the first three formant peaks and bandwidths, pitch, short-time energy, magnitude, and zero crossing rate. The first three formant peaks and bandwidths are estimated using root finding from LPC. Fundamental Frequency is estimated by two different methods: SWIPE, and CREPE. Since these methods have different failure modes their estimates are averaged to produce a final estimate for fundamental frequency. Table 4.1 shows the standard deviation of each of these features on TIMIT to show their reliability across domains. The phonetic transcripts of TIMIT are used

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 4.1: 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.

to look at the consistency of each of these features with phonetic categories. Since we expect the first two formant peak values to be consistent for vowels, we compute the variance of each feature within voiced phoneme categories and report the square root of the average as σ_{phone}. These features are the core of the evaluation described in this thesis, which will focus on demonstrating whether a representation is related to these features. An additional line of work is in showing that a representation is disentangled with respect to these features, since measures of this can be used to show the success of disentanglement in speech representation learning. In Figure 4-1 a spectrogram from TIMIT is visualized alongside the first three formant estimates. The formant estimates are masked during unvoiced segments of speech since the formant estimates are extremely noisy and unreliable without periodic signals. The SWIPE algorithm signals that a frame is likely unvoiced by returning a fundamental frequency of zero in that frame. The first two formants are relatively accurate, but the true second formant is often skipped when the gap between the second and third formants becomes small. The next sections will motivate and describe our approach to using these features as a signal in quantitative and qualitative evaluation.

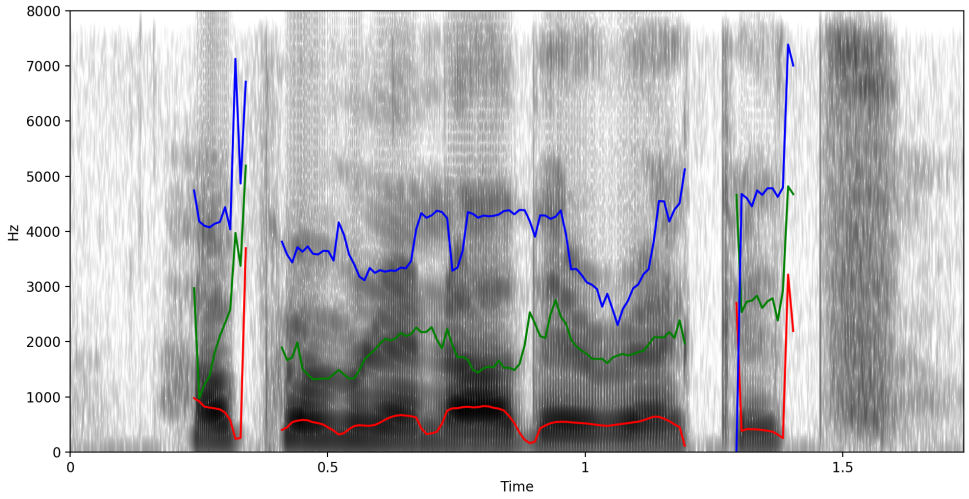

Figure 4-1: Visualization of formant tracking in a spectrogram. The first three formant estimates are visualized in red, green, and blue respectively.

4.1 Acoustic Factors

The first and seconds formants (F1 and F2) are known to be good indicators of tongue position. Since tongue position is a primary differentiator in articulating different phonemes, F1 and F2 have long been tied to predicting vowel category. However, there is obvious overlap between the occurrence of different phonemes on the F1 X F2 plane, Figure 4-2 visualizes this on the TIMIT dataset. There have been several approaches to separating this overlap. One idea is that the F1 X F2 plane is simply insufficient as a feature space, and that human perception uses other acoustic features to define a feature space that separates vowels. Other work has shown that studying the trajectory of each formant across the frames of a vowel can be sufficient in distinguishing vowels. One of the first word recognizers (Davis et al., 1952) used this idea to compare words based on the trajectory of the first three formants. Lastly, the separation might be made possible after adapting to cues about speaker specific warping like fundamental frequency or the average F3 frequency (Harrison, 2013).

We should expect that the representations we wish to evaluate may learn along similar patterns. A representation may learn nonlinear combinations of the features we are interested in and still effectively capture the variation in the data. Represen-

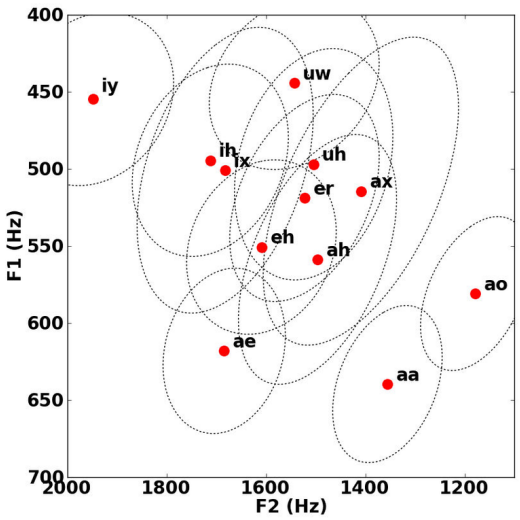

Figure 4-2: Visualization produced in (Weber et al., 2016) that shows approximate clusters for each vowel on the F1 X F2 plane.

tation may encode a trajectory of values into a few parameters and representations may encode both local and global properties of speech. This motivates two specific requirements for the evaluation system and the generative factors used by it. First, generative factors must be able to capture feature trajectories across multiple frames. Second, the generative factors should capture both local properties of speech and global ones. In addition, the methods should be robust to errors in the estimation of the acoustic features. Since the evaluation should apply generally to speech data and not be constrained to specific sub-domains, there should be some mechanism for reducing the impact of noise.

To address these issues we propose to use the statistics of a feature trajectory, the mean and variance, to represent the entire trajectory. We do this at the segment-level to capture features that correlate with the true segment-level identity. At the utterance-level, the feature statistics should capture tendencies of the speaker, or other global properties of the utterance like channel effects or noise conditions. These two sets of factors are referred to as the segment-level and sequence-level factors respectively. Some representations may be predictive of segment-level identity without having any dimension correlate with segment-level factors. If the representation en-

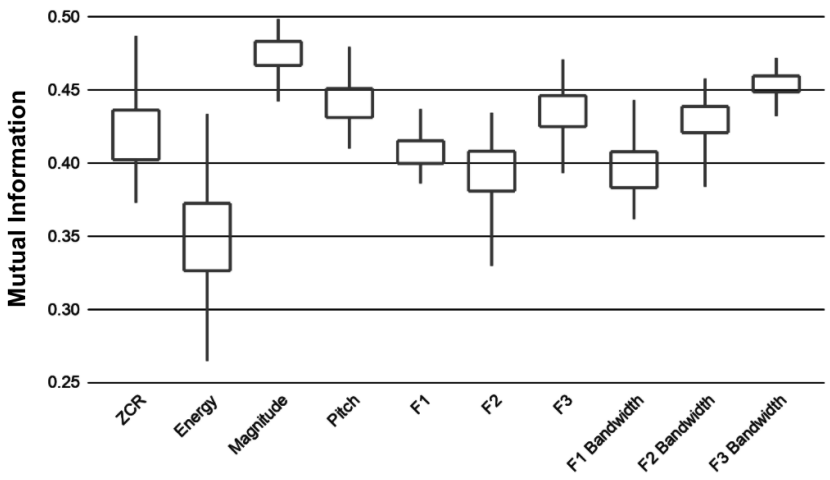

Figure 4-3: Boxplots for the MI between each discretized feature and the mel-spectrogram representation.

codes sequence-level factors then it can encode the trajectory of segment-level factors relative to their global statistics. Because of this we also extract segment-level factors normalized by the global statistics.

In order to reduce the noise of the signal that these features provide and to make the computation of future quantitative metrics easier, we also consider discretizing the factors. We discretize each factor in terms of both the segment mean and the segment standard deviation. The two hyper-parameters for this process are the number of ranges chosen for the segment means and the number of ranges chosen for the segment standard deviations. Each bin corresponds to a range of means and a range of standard deviations, so the product of these hyper-parameters is the total number of bins. To evaluate the efficacy of this approach, we verify that the discrete factors cluster the data similarly for a reasonable range of choices of hyper-parameters. We estimate the mutual information between each extracted segment-level feature and 20 dimensional MFCC feature vectors for many pairs of these hyper-parameters so that the total number of bins varies from 25 to 2500. Figure 4-3 shows the minimum, maximum, and range for the standard deviation of the values reported for each feature. 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.

4.1.1 Limitations

Features like dialect add another interesting layer of complexity since they might not be predictable from either segment-level features or utterance-level features. Some dialectal variation can be explained simply by the presence or absence of certain phonemes of phoneme bigrams. For instance, the presence of /ɹd/ is indicative of North American English as opposed to typical British English, since North American English differs from British English in that it pronounces /ɹ/ before consonants. Other dialectal variation is impossible to predict without more context on the distribution of the language. A dialect may differ in terms of a transformation that applies in some contexts but not all, meaning that no phoneme n-grams are exclusive to one dialect. This illustrates a larger challenge, which is that not all the important variation in speech can be captured using low-level features. As we have discussed, modeling certain types of speaker variation accurately requires aggregating over longer windows of time. Similarly, in order to capture all dialectal variation higher order features must be composed from the linguistic variation and the variation of individual segments. This is a limitation of choosing features by hand in order to define evaluation methods as this thesis does. In order to measure if a representation captures utterance-level variation we must define a process for aggregating over utterances. To do something similar for dialects or other higher order concepts, similarly hand-engineered decisions would need to be made. This puts a limit on what types of variation can be modeled by the methods in this thesis, and so we focus on measuring only the segment-level and utterance-level variation.

4.2 Analysis

For a given speech representation, which encodes either a short segment of speech or the entirety of an utterance, we now have methods for extracting a set of factors that describe the corresponding speech signal. This section will describe concrete quantitative and qualitative analysis that makes use of this signal in order to evaluate a representation.

4.2.1 Quantitative Metrics

(Chen et al., 2018) describes a method for estimating the mutual information between a parameterized posterior latent distribution $p(z|x)$, which is usually Gaussian, and a set of discrete ground truth latent factors v. For probabilistic representations, this can be used to measure how well each axis of a representation captures all or some of the acoustic factors. For point representations, nearest neighbors can be used to cluster the continuous values before computing the mutual information between the discrete factors and the clusters of representations.

Since the mutual information for continuous variables can be difficult to estimate reliably without bias (Ross, 2014), we also propose to use the equal error rate. The equal error rate is a robust analog, for accuracy that attempts to measure the ability of a system to distinguish between discrete categories. By computing a threshold at which the system rejects positive samples at the same rate as it accepts negative samples, equal error rate measures accuracy without including bias in the likelihood of each target category. The equal error rate is a good option for measuring the correspondence between continuous representations and discrete generative factors. By using cosine distance to compute a similarity matrix between pairs of representations, equal error rate can be computed to measure the ability of the latent space to separate each bin of the generative factors.

In order to quantify disentanglement, the mutual information gap (MIG) (Chen et al., 2018) metric can be used. MIG evaluates disentanglement as the degree to which each generative factor is captured by only one dimension of the representation.

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 4.2: 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.

Formally, for K factors v_k, the entropy of each factor $H(v_k)$, mutual information over n samples I_n, and $j^{(k)} = argmax_j I_n(z_j; v_k)$ the mutual information gap is

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H(v_k)}(I_n(z_{j^{(k)}}; v_k) - \max_{j \neq j^{(k)}} I_n(z_j; v_k)) \tag{4.1}$$

4.2.2 Generated Data

Not all of the variation that exists in speech data will be captured by any given representations. Representations, in general, will learn patterns within the data that allow observations to be encoded more concisely. For instance, a method might learn gender specific average formant centers, particularly for higher formants, in order to avoid encoding every point along the trajectory. Instead, we should expect that the representation's estimate for each gender's average formant center would be baked into the neural network approximating the generative distribution. One path for analysis is to analyze the reconstructed data produced by the generative model. The relationship between a representation and the reconstructed data will do a better job at measuring patterns that may be hidden within the models themselves. Acoustic features can be useful in this case too since we can use them to describe the properties of the reconstructed data.

Previous work has attempted to verify that changes in a latent dimension correspond to specific changes in the output of the generative model (Hsu et al., 2018). 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 4.2 reproduces Table 4.1 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.

Chapter 5

Analyzing Methods for Representing Speech

This section describes the methods used to test the evaluation suite. The factorized hierarchical variational autoencoder (FHVAE) (Hsu et al., 2017b; Hsu and Glass, 2018) 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 (Garofolo et al., 1992), with 240 utterances from 24 speakers reserved for evaluation.

5.1 Visualizing Disentanglement

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 5-1, 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

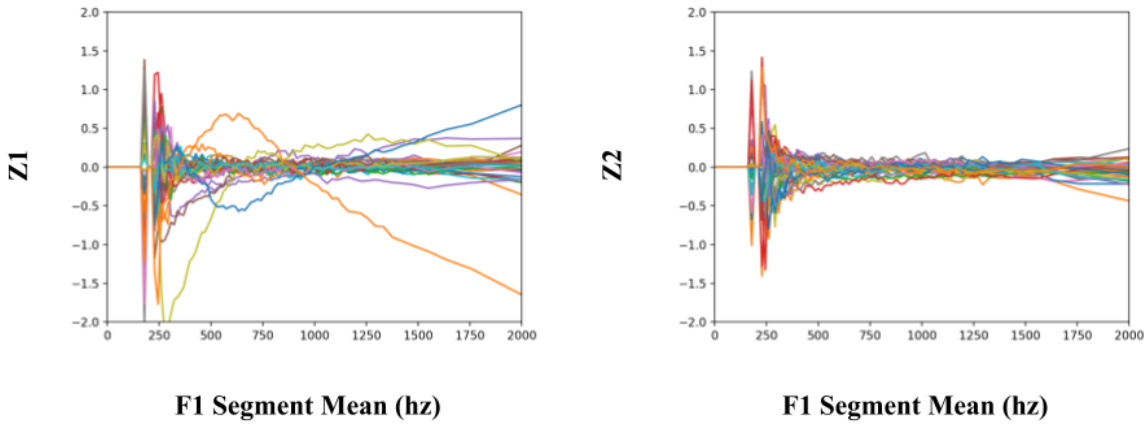

Figure 5-1: Both z_1 and z_2 are compared to F1, several dimensions of z_1 take extreme values due to change in F1 and no such trend is seen for z_2. Spikes at low frequencies are due to small numbers of samples.

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.

Since z_1 and z_2 are encodings of information from different timescales we are also interested if they have different sensitivities to features extracted at different timescales. For this analysis, we extract features for the 200ms segments corresponding exactly to each representation and segments with the same centers but corresponding to 2000ms instead. Figure 5-2 shows the trends between z_1 and z_2 and pitch, with the shorter timescale on top and the longer timescale on bottom. We see that the dimensions z_1 have similar relationships with the pitch at both timescales. This might indicate that the variation in pitch is small enough that the mean pitch of a 200ms segment is a good predictor of the mean pitch of the 2000ms segment with the same center. This can not be ruled out unless the variation across timescales is better quantified, meaning that this visualization can't say much about a trend disappearing at a longer timescale. However, for z_2, we can see evidence that a trend

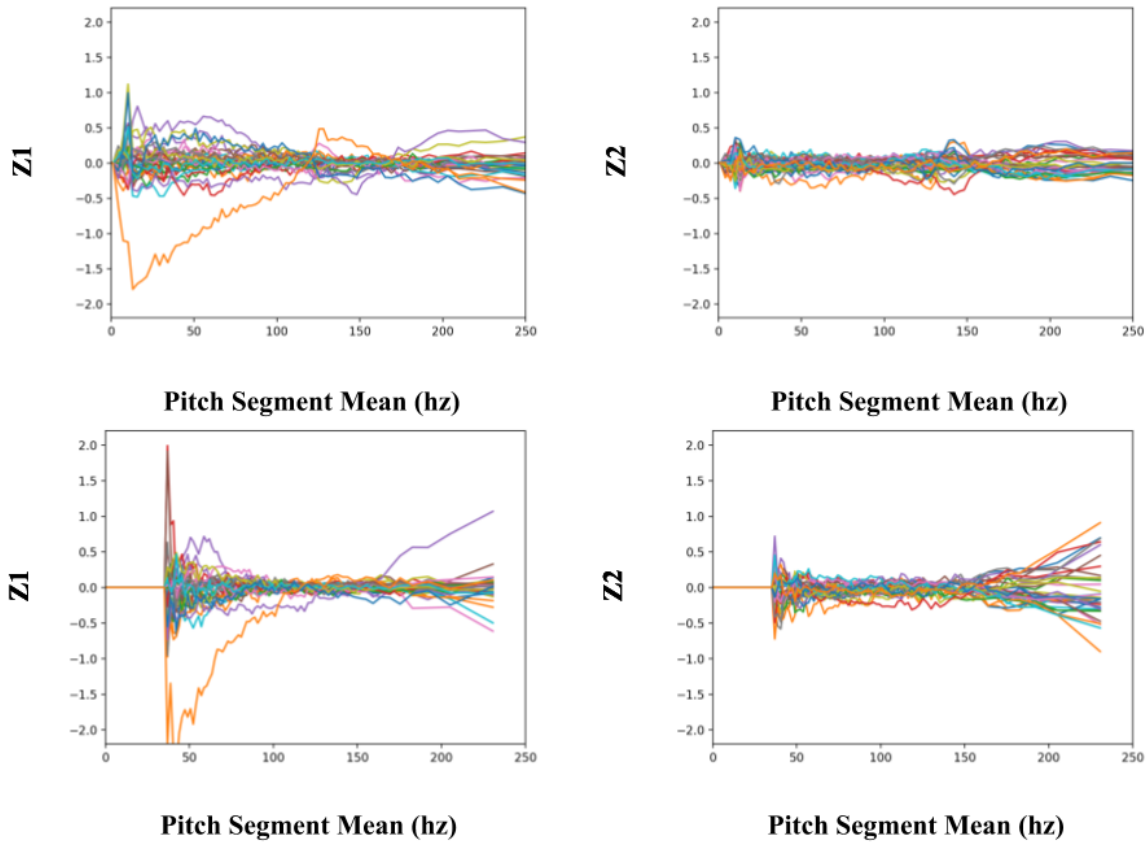

Figure 5-2: On the top z_1 and z_2 are compared to pitch at segments of 200ms, on the bottom pitch is aggregated instead over 2000ms segments.

appears only at a longer timescale. This is consistent with our understanding of z_2. Since it encodes information about utterance-level variation it should be sensitive to utterance-level factors like gender, which is closely related to pitch.

5.2 Characterizing Sensitivity of Latent Dimensions

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 5-3 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

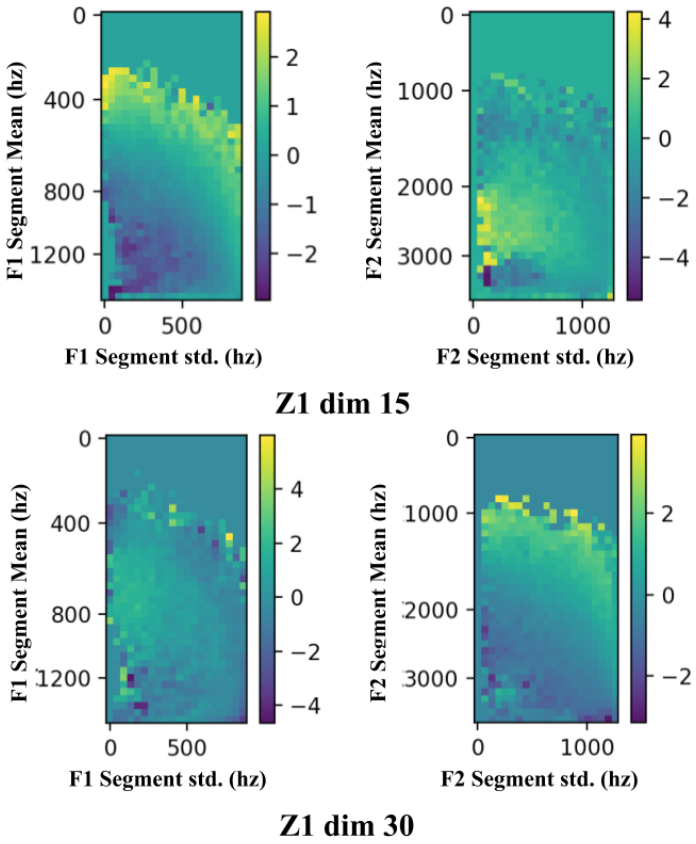

Figure 5-3: Two dimensions of z_1 are plotted with respect to F1 and F2. Each bin corresponds to a range of segment means and segment standard deviations.

the pixels display the average intensity of the 15^{th} dimension for samples from the respective partition. This process is repeated in each 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 5-4 we produce analogous plots that measure the frequency of the corresponding phonetic

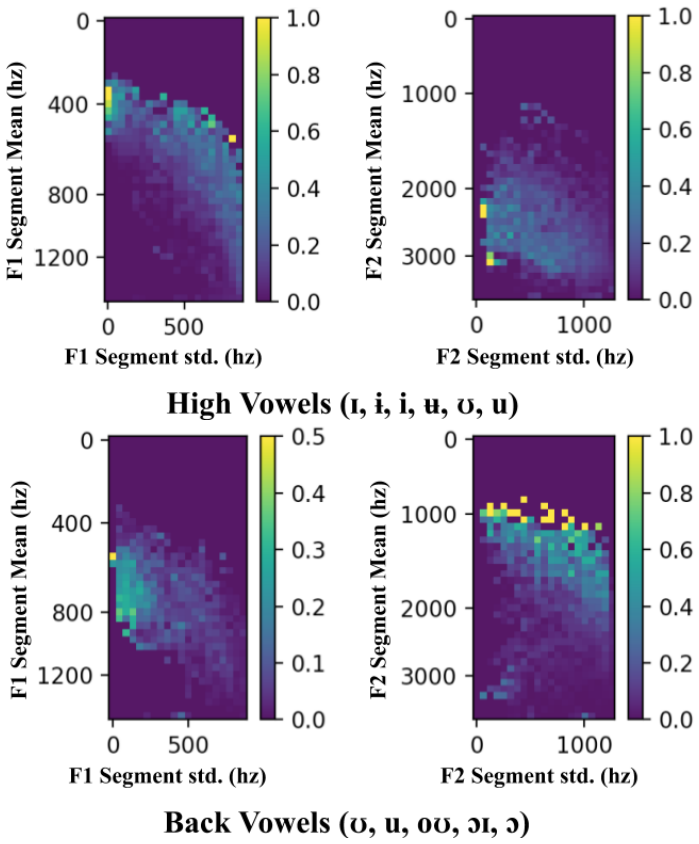

Figure 5-4: The frequency of two phonetic categories are plotted with respect to F1 and F2. Each bin corresponds to a range of segment means and segment standard deviations. Note the correspondence with Figure 5-3.

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 than real valued intensities.

Lastly, we reproduce this analysis through another lens. The two latent dimensions that were identified to be sensitive to particular vowel categories are plotted with respect to only the F1 and F2 segment means. In Figure 5-5, we observe the same patterns as before, the 15^{th} dimension is sensitive to segments with low F1 the 30^{th} is sensitive to segments with low F2. These plots aggregate over the segment variances of each feature and lose the ability to distinguish along that axis. However, we can see that the descriptive ability is similar and this is another option for identifying relationships between latent dimensions and phonetic categories.

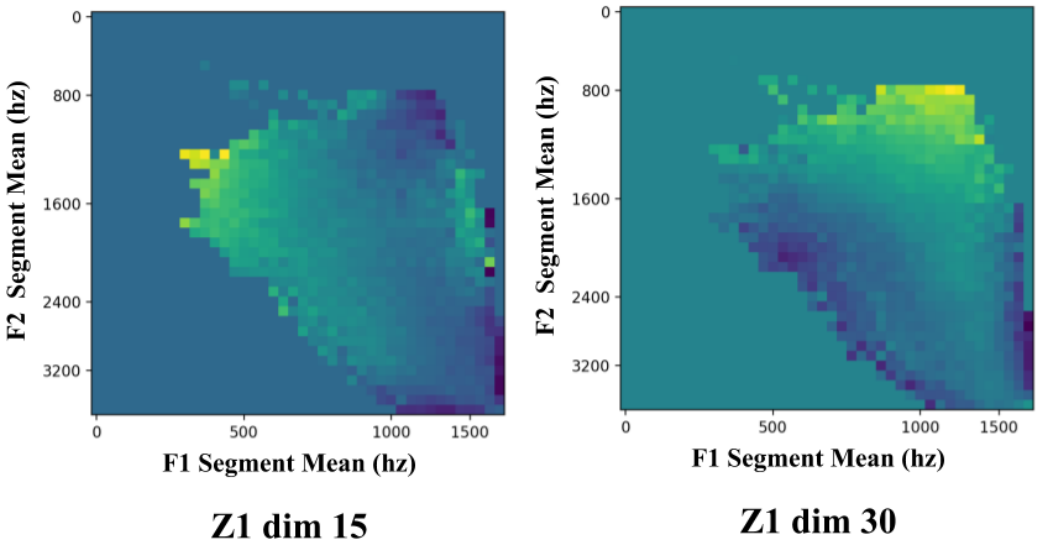

Figure 5-5: Two dimensions of z_1 are plotted on the F1 X F2 plane with respect to the segment-means of each feature.

Chapter 6

Evaluation Suite

This chapter describes an evaluation suite designed to make the analysis described in previous chapters available for a variety of representation learning research. The tool is built in python as a set of analysis scripts that share a processing and evaluation backend. Users are expected to have their own training and extraction pipelines to produce representations extracted across every sound file. The suite natively supports TIMIT and Librispeech and users can run analysis on these datasets without needing to specify the structure of the dataset. Running analysis on other datasets requires that the user specify the directory structure of the dataset so that the tool can match the provided representations to the raw speech signals. Transcriptions or other forms of supervision can be made available to the toolkit. Since the format of the additional data must be specified by the user and there are no general patterns for how this might be done, the tool requires that additional supervision be specified programmatically. The API is designed so that this process is simple and examples are given in the package.

When an evaluation script is run by the user the tool will first extract all the required acoustic features. Since this is relatively time-consuming, $\sim$20 minutes for TIMIT on a single machine, the tool will save the results into the hdf5 format for future use. In this way, future analysis on the same dataset is much faster. The tool extracts formant and formant bandwidths, energy, magnitude, zero crossing rate, and two estimates of pitch. Analysis scripts read the provided representations and deter-

mine the slices of the data that each representation corresponds to. By default, it is assumed that the representations encode the data with a constant window and step size and this is used to compute the appropriate segments. However, the user can specify the window and step size that was used to generate their data if necessary. The tool can also be applied to representations that encode variable-length segments, for instance, if the tool has been trained to detect silence and to encode the variable-length snippets between adjacent silences. In this case, the user can specify the boundaries of each representation relative to the total length of each utterance. Examples are given for each of these cases. Additional acoustic features can be added by the user just as they can specify the source of additional transcriptions or other supervision. The processing pipeline is implemented as a set of dependent feature extractors that are connected at runtime, so the user can easily make use of our acoustic processing tools and avoid re-implementation.

Implementations for mutual information, mutual information gap, and equal error rate are provided as quantitative methods. Each of these is computed with respect to discretized acoustic factors as described in section 4.1. By default, we use 10 bins for the the segment-means and 5 for the segment-variances. These hyper-parameters can be set by the user and as demonstrated the metric values should be robust within a relatively wide range of values. Three visualization tools are also included to make use of the methods described in the previous chapter. The first plots the trend between each latent dimension and a single discrete acoustic factor, which can be a segment-level or an utterance-level factor. Examples of this are given in section 5-1. By plotting the average value of each latent dimension within each bin, disentanglement can be visualized. The same plot can be produced for data generated by the user using a generative model and the provided representations. In this case, the tool will estimate feature values for each reconstructed waveform and produce the same plot. Last, the visualization of the F1 X F2 plane is included as an easy method for visualizing whether a representation is sensitive to vowel identity and an example of this can be seen in 5-5.

Chapter 7

Conclusion

7.1 Summary of Contributions

This thesis discusses methods for using acoustic processing to aid in the evaluation of unsupervised representation learning. We review research across speech processing in order to motivate specific analysis techniques that are grounded in linguistically significant features. This work addresses the difficulty in rigorously evaluating the significance of methods in representation learning by providing an additional source of structure.

Chapter 4 discussed how acoustic processing could be used to track variation in short segments of speech and across utterances. We introduced a methodology for extracting continuous and discrete features that provide an interpretable signal about speech. Several quantitative and qualitative methods were defined to make use of this signal. In addition, a method for evaluating the generative decoders of deep generative models was described.

Chapter 5 applied these methods to produce qualitative and quantitative analysis for previous work on unsupervised representation learning. We validate previous understanding of FHVAE using these methods and perform novel analysis of the representations it produces. The evaluation makes use of generative factors extracted at multiple scales, to consider both the segment-level and utterance-level variation.

Chapter 6 introduces a general-purpose evaluation suite for representations of

speech. This toolkit is open-sourced on Github with the purpose of enabling benchmarked analysis grounded in acoustic processing. The hope is to enable comparison of methods in disentangled representation learning in the complex and real-world domain of speech.

7.2 Future Work

The work of this thesis enables a variety of future research into grounded validation of existing representation learning methods. Future work could produce ablation studies using the criteria from this thesis that use acoustic features. Reproducing previous research through the lens of this thesis would provide credibility to the approach.

Future work should attempt to address specifically how error in estimating acoustic features affects the results of these methods. While estimation error is addressed, its impact on the produced evaluation results is not measured directly.

Another path of research is to extend the methods in this thesis to capture high order linguistic content. This thesis is focused on phonetic information and so acoustic processing and modeling are given specific treatment. Thus, the evaluation suite is best suited for representations encoding lower-level phonetic units. However, this type of analysis is just as suited for encodings of word-like units and long-term dependencies like prosody. To evaluate a representation or a generative model's ability to capture this information, predictive features could be isolated and incorporated into the framework presented here.

Bibliography

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.

Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL `http://arxiv.org/abs/1206.5538`.

G. Cai, Y. Wang, and L. He. Learning smooth representation for unsupervised domain adaptation. *CoRR*, abs/1905.10748, 2019. URL `http://arxiv.org/abs/1905.10748`.

A. Camacho and J. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124:1638–52, 10 2008. doi: 10.1121/1.2951592.

T. Cemgil, S. Ghaisas, K. Dvijotham, and P. Kohli. Adversarially robust representations with smooth encoders. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1gfFaEYDS`.

R. T. Chen, X. Li, R. B. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *CoRR*, abs/1802.04942, 2018. URL `http://arxiv.org/abs/1802.04942`.

K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952. doi: 10.1121/1.1906946. URL `https://doi.org/10.1121/1.1906946`.

A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002. doi: 10.1121/1.1458024. URL `https://doi.org/10.1121/1.1458024`.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

Y. Dissen, J. Goldberger, and J. Keshet. Formant estimation and tracking: A deep learning approach. *The Journal of the Acoustical Society of America*, 145(2):642–653, 2019. doi: 10.1121/1.5088048. URL `https://doi.org/10.1121/1.5088048`.

S. A. Fattah, W. P. Zhu, and M. O. Ahmad. A cepstral domain algorithm for formant frequency estimation from noise-corrupted speech. In *2008 International Conference on Neural Networks and Signal Processing*, pages 114–119, 2008.

D. Gagouri and A. Kammounand Ahmed Hamida. A comparative study of formant frequencies estimation techniques. 01 2006.

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.

J. S. Garofolo, D. Graff, D. Paul, and D. Pallett. CSR-I (WSJ0) Other, 2016. URL `https://doi.org/10.7910/DVN/ZVU9HF`.

P. Harrison. *Making accurate formant measurements: an empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. PhD thesis, University of York, 2013.

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. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

W. Hsu and J. Glass. Scalable factorized hierarchical variational autoencoder training, 2018.

W. Hsu, Y. Zhang, and J. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data, 2017a.

W. Hsu, Y. Zhang, and J. R. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *CoRR*, abs/1709.07902, 2017b. URL `http://arxiv.org/abs/1709.07902`.

W. Hsu, Y. Zhang, R. J. 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*, abs/1810.07217, 2018. URL `http://arxiv.org/abs/1810.07217`.

M. Jalil, F. A. Butt, and A. Malik. Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In *2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, pages 208–212, 2013.

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)*, pages 1614–1618, 2017.

S. Khurana, S. R. 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)*, pages 6540–6544, 2019.

H. Kim and A. Mnih. Disentangling by factorising, 2018.

J. W. Kim, J. Salamon, P. Li, and J. P. Bello. Crepe: A convolutional representation for pitch estimation, 2018.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2013.

Y. Li and S. Mandt. Disentangled sequential autoencoder, 2018.

F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *CoRR*, abs/1811.12359, 2018. URL `http://arxiv.org/abs/1811.12359`.

A. Makhzani and B. Frey. k-sparse autoencoders, 2013.

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

T. M. Nearey. Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, 85(5):2088–2113, 1989. doi: 10.1121/1.397861. URL `https://doi.org/10.1121/1.397861`.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

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*, AVSS '09, page 296–301, USA, 2009. IEEE Computer Society. ISBN 9780769537184. doi: 10.1109/AVSS.2009.58. URL `https://doi.org/10.1109/AVSS.2009.58`.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19 – 41, 2000. ISSN 1051-2004. doi: https://doi.org/10.1006/dspr.1999.0361. URL `http://www.sciencedirect.com/science/article/pii/S1051200499903615`.

B. C. Ross. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):1–5, 02 2014. doi: 10.1371/journal.pone.0087357. URL `https://doi.org/10.1371/journal.pone.0087357`.

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, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1449.

S. Shon, W. Hsu, and J. Glass. Unsupervised representation learning of speech for dialect identification, 2018.

R. C. Snell and F. Milinazzo. Formant location from lpc analysis data. *IEEE Transactions on Speech and Audio Processing*, 1(2):129–134, 1993.

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937. doi: 10.1121/1.1915893. URL https://doi.org/10.1121/1.1915893.

D. Talkin. A robust algorithm for pitch tracking (rapt). 2005.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. 2016. URL http://arxiv.org/abs/1609.03499. cite arxiv:1609.03499.

P. Weber, L. Bai, M. Russell, P. Jancovic, and S. Houghton. Interpretation of low dimensional neural network bottleneck features in terms of human perception and production. 09 2016. doi: 10.21437/Interspeech.2016-124.