INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Vowel Normalisation in Latent Space for Sociolinguistics

James Burridge

School of Mathematics and Physics, University of Portsmouth, United Kingdom

james.burridge@port.ac.uk, james.burridge@gmail.com

Abstract

To study variations in vowel sounds between different sociolinguistic groups, sounds must be normalized to minimize variations caused by physical factors. The Lobanov method, for example, standardizes formant distributions by speaker. Since formants are often difficult to measure, and offer only a partial description of sounds, a robust and reproducible normalisation method based on the whole spectrum would be useful. One candidate is speaker-level standardization in the latent space of a variational auto-encoder, trained on a large sample of vowel spectra. We show that whole spectrum transformations induced by latent normalisation shift formants similarly to direct formant normalisation. We also show that formant-based normalisation procedures can be used to induce whole-spectrum transformations via latent space.

Index Terms: normalisation, formants, vowels, dialects, sociolinguistics

1. Introduction

Variations in the way that people use language based on where they are from, or on their social characteristics, have long fascinated both professional linguists and the wider population. These variations can occur at many levels, including lexicon, syntax, prosody and the inventory of individual phones or phonemes used by each speaker [1, 2]. At the level of phonemes, vowel sounds, which are the primary carriers of linguistic information in connected speech [3, 4], exhibit a particularly wide range of variations both in terms of the inventory used by each speaker, and how this inventory is deployed within different words [5]. For example Figure 1, derived from the Survey of English Dialects [6], gives an approximation of how the fraction of speakers who use a similar vowel sound in the words "foot" and "strut" varied geographically in England amongst speakers born near the start of the 20th century. The same pattern, marked by a distinctive east-west "isogloss" persists to this day. The existence of these patterns mean that linguistic variations between speakers can be used to draw inferences about their geographical origins and social group [7]. Beyond academic interest, such inferences have applications in combating linguistic profiling and in forensics [8].

1.1. Speaker normalisation for sociolinguistics

In order to use sound recordings to study variations in the sounds used by different speakers, an essential first step is to normalize measurements extracted from them to reduce variations caused by physical differences between the vocal apparatuses of the speakers [9]. In the case of vowels, which are resonant sounds made without significant restriction of airflow through the vocal articulators [10], the most common measurements of interest are formants [11, 12]. Although there is more than one definition of these quantities [13, 14], they are commonly understood to be resonant frequencies of the supraglottal vocal tract, typically measured as the locations of peaks of the spectral envelope, and labelled $F_1, F_2, F_3, \ldots$ in order of increasing frequency. They may be calculated from linear prediction coefficients estimated from a short window of the speech signal [12]. The salience of formants to the measurement of vowels derives in part from their close relationship to the traditional linguistic features of tongue height and backness [15] which determine position on the vowel quadrilateral, used near universally as a means to describe vowel inventories in world languages [16, 17]. Formants are widely used as a basis for speaker normalisation because of their connection to the size and shape of the vocal tract. For a given configuration of the vocal articulators (tongue, lips, teeth) a speaker with a larger vocal tract will produce lower formants [18].

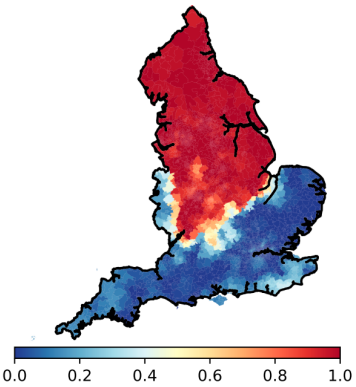

Figure 1: *Fraction of speakers for whom the vowels in "foot" and "strut" rhyme. Data taken from the Survey of English Dialects.*

Much research has been carried out to find the optimal formant transformations to remove sound variations caused by physical differences without also suppressing socio-linguistic variations [19]. In this paper we will consider the simple but effective Lobanov method, where each formant value is standardized for each speaker, s, using the mean and variance of its value over a set of utterances representative of the range of sounds uttered by that speaker. That is

$$F_i^\dagger(s) = \frac{F_i - \mu_i(s)}{\sigma_i(s)} \tag{1}$$

where $F_i^\dagger$ is a single normalized measurement of the ith for-

10.21437/Interspeech.2023-1704

mant, and $\mu_i(s), \sigma_i(s)$ are its mean and standard deviation over the representative sounds of speaker s. Formants normalized in this way can be transformed back to typical Hertz values using a reference mean and standard deviation μ_i and σ_i calculated from the population as a whole. The final adjusted formant is

$$\tilde{F}_i(s) = \mu_i + \sigma_i F_i^{\dagger}(s). \tag{2}$$

The mean and variance is therefore the same for all speakers, with sociolinguistic variations assumed to be captured by other characteristics of the distributions of the $\tilde{F}_i$. We refer to equation (2) as *standard* Lobanov normalisation.

1.2. Problems with formants

The widespread use of formants as speech features is testament to their usefulness in applications, including dialect comparison [20, 21], measuring linguistic variation and change [1], speech synthesis [22, 23] and the analysis of vocal tract shapes and configurations [24]. However, they give only a partial description of sounds and, more importantly, there are problems with their automatic measurement which may not be fully resolvable [25]. Their estimation involves the use of algorithms with parameters which must be adjusted to achieve reasonable results, and if one formant is incorrectly identified this affects all higher formants [14]. Further, it may not be reasonable to view all vowel sounds as composed of a small number of simple resonances. In summary, when they work, it is hard to beat formants as a quantitative measure of vowel sounds in terms of efficiency and explanatory power. However, they are neither always measurable nor always meaningful, and they do not offer a complete description of the perceptual characteristics of vowels [26]. When normalizing speakers for sociolinguistic analysis, we would like a technique which can transform the whole spectrum, and which possesses the positive characteristics of methods based on formants, but does not necessitate their measurement.

1.3. Labonov normalisation in the latent space of a Variational Autoencoder

An unsupervised learning technique which has found many applications in speech technology is the variational auto-encoder (VAE) [27, 28, 29]. The VAE is a probabilistic generative model including latent variables which encode key features of high dimensional data such as audio. The VAE has been used for various forms of speaker normalisation [30, 31], for example to improve automatic speech recognition using small amounts of audio for each speaker [30]. Our focus is on vowels, where substantial data is available for each speaker, and a simple and reproducible method is needed to normalize differences in vowel sounds for sociolinguistic analysis. We note that recent work has shown that formants can be successfully encoded by latent VAE features [32], indicating the potential of the VAE to improve on formant normalisation, whilst retaining its useful characteristics.

We briefly review essential principles of the VAE. It is a joint probability model $p(\mathbf{x}, \mathbf{z})$ where $\mathbf{x} \in \mathbb{R}^N$ is a high dimensional random vector representing observable data, and $\mathbf{z} \in \mathbb{R}^M$ is a lower dimensional "latent" random vector which captures the salient features of $\mathbf{x}$. In the case of vowels, we view $\mathbf{x}$ as the spectrum, and we might expect the corresponding $\mathbf{z}$ vectors to encode features which are closely related to formants. In this paper we use a Gaussian VAE, for which the distribution of $\mathbf{x}$ conditional on $\mathbf{z}$ is defined to be

$$p_\theta(\mathbf{x}|\mathbf{z}) \triangleq \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{z}), \sigma^2 I) \tag{3}$$

where I is the $N \times N$ identity matrix and $\boldsymbol{\mu}_\theta(\mathbf{z})$, known as the *decoder*, is a neural network with parameter vector θ, which maps latent vectors to points in $\mathbb{R}^N$. The variance σ^2 is a constant whose relevance we describe below. The prior distribution of $\mathbf{z}$ is defined to be standard multivariate normal

$$p(\mathbf{z}) \triangleq \mathcal{N}(\mathbf{z}|0, I) \tag{4}$$

where I is the $M \times M$ identity matrix. The intractability of the marginal likelihood $p_\theta(\mathbf{x})$ and the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ is dealt with by introducing a variational approximation [33] to the posterior

$$p_\theta(\mathbf{z}|\mathbf{x}) \approx q_\phi(\mathbf{z}|\mathbf{x}) \triangleq \mathcal{N}(\mathbf{z}, \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}\{\boldsymbol{\sigma}^2_\phi(\mathbf{x})\}) \tag{5}$$

where the parameter vectors $\boldsymbol{\mu}_\phi(\mathbf{x}) \in \mathbb{R}^M$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{x}) \in \mathbb{R}^M$ are the outputs of an *encoder* neural network with parameter vector ϕ. The VAE is trained by maximizing a lower-bound of $\ln p_\theta(\mathbf{x})$ averaged over the data. This "evidence lower bound" (ELBO) is defined

$$\mathrm{ELBO}(\mathbf{x}) \triangleq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}(\ln p_\theta(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{6}$$

where D_{KL} denotes the Kullback-Leibler divergence. The first term on the right in (6) is

$$-M \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \int \|\boldsymbol{\mu}_\theta(\mathbf{z}) - \mathbf{x}\|^2 q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{z} \tag{7}$$

which measures the reconstruction error of the decoder $\boldsymbol{\mu}_\theta(\mathbf{z})$ averaged over the distribution of latent vectors generated by the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ for given input $\mathbf{x}$. From this we see that smaller values of σ result in a higher penality for reconstruction errors. The second term on the right in (6) measures the extent to which the posterior distribution departs from the standard normal prior, and acts as a regularization term which encourages the marginal distribution of latent vectors generated by the empirical data to be as close to standard, uncorrelated multivariate normal as possible.

The aim of this paper is first to explore the effect of transferring the process of Lobanov normalisation to the latent features, by considering its effect on decoded spectra. Having considered direct latent space normalisation, we investigate latent transformations which change the whole spectrum, based only changes in formants. Finally we consider the extent to which formants are disentangled in latent space.

It is important that normalisation methods used for sociolinguistic analysis are simple, transparent and reproducible. Since simpler techniques are preferable in this regard, we work with the most basic kind of VAE. The contribution of our paper is to illustrate the potential advantages of using probabilistic latent variables models as an alternative to formant based methods for performing speaker normalisation for sociolinguistic analysis.

2. Methodology

For simplicity in this paper we take $\mathbf{x}$ to be order 16 LPC smoothed log power spectra [12] evaluated at 100 frequencies in the interval $[0, 4000]$ Hz, equally spaced on the mel scale [34]. This is a wide enough frequency range to contain F_1, F_2 and F_3 for most speakers [20]. We computed spectra at the midpoint of each vowel in the TIMIT database [35], using 512 frames (32ms at 16 kHz sample rate, Hanning window applied), and normalize so that the average power over all frequencies is zero for each spectrum. Our experience with formants suggests that each such spectrogram should be describable using

a small number of features, perhaps six if we believe that formant peak locations and heights vary independently, so we set $M = 6$. We found that larger M values lead to redundant latent directions (zero variance). Our encoder and decoder are simple multi-layer networks with hidden layer sizes $[100, 100, 50]$ and $[50, 100, 100]$ respectively. Experimentation with σ revealed that $\sigma^2 = 0.1$ produced reconstructed spectra with minimal distortion (such as spurious extra formant peaks) whilst maintaining a near-normal marginal posterior distribution. We implemented the VAE using PyTorch [36] and trained using the Adam optimizer for 80 epochs with batch size 128 and learning rate 10^{-3}. Reduction in the overall loss was negligible beyond this point.

For each spectrum in our dataset we computed the first three peak locations (formants) and also its latent features via the encoder. We then fit K-nearest-neighbours regressors $\hat{F}_i(\mathbf{z})$ [37] with $K = 100$ and Gaussian weights (bandwidth 0.5, chosen to minimize test prediction error) to predict formant values from latent position, obtaining root mean squared test errors of 38Hz, 145Hz and 299Hz for predictions of F_1, F_2, F_3. The regression functions allow us to approximate formant gradients $\nabla\hat{F}_i(\mathbf{z})$ in latent space using finite differencing. In this paper, for simplicity, we focus on normalisation of the first two formants. We performed the three experiments listed below, with results reported in section 3. Python implementations of core functions used are publicly available [38].

2.1. Lobanov normalisation in latent space

To investigate the effect of applying a standard speaker normalisation method to latent variables we standardized the latent features of each speaker as follows

$$\tilde{\mathbf{z}} = \frac{\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}(s)}{\sigma_{\mathbf{z}}(s)} \tag{8}$$

where $\boldsymbol{\mu}_{\mathbf{z}}(s)$ and $\sigma^2_{\mathbf{z}}(s)$ are the mean and variance of the latent features for speaker s. We then computed the transformed spectrogram, $\tilde{\mathbf{x}}$, for each vowel utterance of each speaker using the decoder

$$\tilde{\mathbf{x}} = \boldsymbol{\mu}_\theta(\tilde{\mathbf{z}}). \tag{9}$$

We refer to this as *direct* latent space normalisation. Finally we computed the spectral peaks of these transformed spectrograms, and compared them to the corresponding Lobanov normalized formants computed using equation (2).

2.2. Latent space transformation via gradient ascent

To induce whole spectrogram transformations from Lobanov formant shifts we performed the following gradient ascent procedure using our regression functions $\{\hat{F}_i(\mathbf{z})\}_{i=1}^2$. For each spectrogram, $\mathbf{x}$, we first computed the Lobanov normalized formants $\{\tilde{F}_i\}_{i=1}^3$ according to equation (2). We then iterated the following transformation in latent space, starting from the latent position, $\boldsymbol{\mu}_\phi(\mathbf{x})$, generated by the encoder

$$\mathbf{z}_{n+1} = \mathbf{z}_n + \alpha \sum_{i=1}^{2} (\tilde{F}_i - \hat{F}_i(\mathbf{z}_n)) \nabla\hat{F}_i(\mathbf{z}_n) \tag{10}$$

with $\alpha = 10^{-5}$, until $|\tilde{F}_i - \hat{F}_i(\mathbf{z}_n)|$ reached a specified tolerance (15Hz) or the number of iterations exceeded a given maximum. We then computed the transformed spectrograms using the decoder. We call this *induced* latent space normalisation.

2.3. Measuring disentanglement

We measured the disentanglement of formants in latent space using two methods. Two features are disentangled in latent space if the normal vectors to the hyper surfaces on which they are constant are orthogonal. If F_i and F_j are disentangled

$$\nabla F_i(\mathbf{z}).\nabla F_j(\mathbf{z}) = 0 \tag{11}$$

for all $\mathbf{z}$. To quantify the extend of disentanglement we computed the empirical distribution of

$$\frac{\nabla\hat{F}_i(\mathbf{z}).\nabla\hat{F}_j(\mathbf{z})}{|\nabla\hat{F}_i(\mathbf{z})||\nabla\hat{F}_j(\mathbf{z})|} \tag{12}$$

over a large random sample of latent points. We then compared this average to the distribution obtained by replacing formant gradients with normal random vectors.

As a second measure of disentanglement, for each nearest neighbour pair of latent vectors $\mathbf{z}_1, \mathbf{z}_2$, corresponding to real spectrograms $\mathbf{x}_1, \mathbf{x}_2$, we computed the vector

$$|F_i(\mathbf{x}_1) - F_i(\mathbf{x}_2)|\mathbf{u}_{12} \tag{13}$$

where $\mathbf{u}_{12} = (\mathbf{z}_1 - \mathbf{z}_2)/(|\mathbf{z}_1||\mathbf{z}_2|)$ is a unit vector in the direction $\mathbf{z}_1 - \mathbf{z}_2$. We then calculated principle components for the set of such vectors for each formant. The cumulative variance explained by the principle components then gives a measure of the dimension of the latent subspace within which each formant varies.

3. Results

The relationship between standard Lobanov formant shifts, and shifts produced by direct latent space normalisation, is illustrated in Figure 2. This shows scatter plots of formant shifts generated for each spectrogram by the two procedures, and we see that they are approximately proportional (linear correlations 50% and 25% respectively for F_1 and F_2 shifts). The mean absolute shifts in F_1 and F_2 caused by Lobanov normalisation are 46.6Hz and 136.8 Hz. To quantify the difference in shifts produced by the two procedures, we let F_i^* denote the value of F_i after transforming latent variables, and define the difference between this an the standard Lobanov normalised formant $\tilde{F}_i$ for a single spectrum, to be

$$\Delta F_i = F_i^* - \tilde{F}_i. \tag{14}$$

Table 1 summarises the statistics of the magnitudes of these differences. Figure 3 shows an example of a spectrogram shifted by this method. The overall structure remains intact with the same number of peaks, with locations close to those computed via direct Lobanov normalisation. Figures 2 and 3 indicate that formant locations have been encoded in the latent feature space, and that speaker normalisation in latent space produces shifts that are closely related to those obtained using direct formant normalisation. However, we emphasise that the relationship between standard Lobanov and latent shifts is quite noisy, as quantified by table 1, and that Figure 3 shows an example where formant and latent shifts are closer than average.

In cases where formants can be reliably measured, the latent variable model provides a method to induce a transformation of the whole spectrogram via gradient ascent (10), using standard Lobanov shifts as targets. A benefit of this approach is that information about the sound, beyond simple formant values, is also present after transformation. Figure 4 shows distributions

Table 1: *50th and 90th centile absolute differences (Hz) in formant shifts produced by standard Lobanov normalisation and (1) direct latent normalisation, (2) induced latent normalisation with transformed formants computed by peak detection, (3) induced latent normalisation with formants estimated by prediction models* $\hat{F}_i(\mathbf{z})$.

Method	$\lvert\Delta F_1\rvert_{50}$	$\lvert\Delta F_1\rvert_{90}$	$\lvert\Delta F_2\rvert_{50}$	$\lvert\Delta F_2\rvert_{90}$
1.Direct Latent	39.9	94.1	117.5	287.8
2.Induced (Peak)	12.1	48.7	29.9	131.9
3.Induced (Model)	9.6	33.4	2.8	11.4

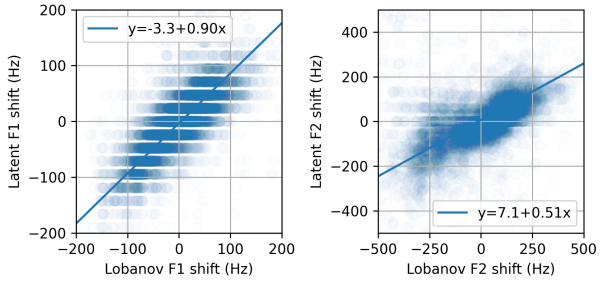

Figure 2: *Scatter plots of formant shifts produced by standard Lobanov normalisation (equation (2)) and by direct latent space normalisation for* F_1 *and* F_2*. Lines show linear fits to the data.*

of ΔF_i (equation (14)) which in this context may be viewed as errors in the final result of the gradient ascent, for a random sample of "significant" shifts ($|\tilde{F}_1 - F_1| > 50$Hz and ($|\tilde{F}_2 - F_2| > 100$Hz). Error statistics are summarised in Table 1.

Finally we consider the extent to which formants are disentangled in latent space. The mean of the normalized dot product in equation (12) over a random sample of 10^3 spectra is only marginally smaller than the same measure where the gradients are replaced with random vectors (0.34 vs. 0.37 for F_1 and F_2), indicating that formants are not disentangled. Our measure of the dimension of the latent sub-spaces within which each formant varies, computed by principle component analysis of the set of vectors defined by equation (13) leads us to similar conclusions. A four dimensional subspace is required to explain 85% of the variation in F_1 and a five dimensional space is needed to explain 87% of the variation in F_2. While disentanglement is not essential for our normalization methods to be effective, recent work [32] suggests that it may be achievable using higher dimensional latent spaces.

4. Conclusions

To analyse sociolinguistic variations in vowel sounds it is essential to normalise speech features [9]. The standard method for achieving this is formant normalisation. One well known example is Labonov normalisation, but many other formant based methods exist [19]. There are two weaknesses with the formant based approach. First, formants are not always a measurable or meaningful way to characterise sounds. Second, they offer only a partial description of the spectrum, and there may be many other interesting features which should also be normalized and used to study language variation. A normalisation method which avoided formants, making it applicable to any sound spectrum, would have the dual advantage of broader applicability, and the normalisation of all features simultaneously. In this paper we have investigated the possibility of using a latent variable model [27] to perform normalisation. We have shown that simple latent standardization at the level of individual speakers produces whole spectrum transformations which generate formant shifts consistent with the standard Lobanov method. We have also shown that spectra can be normalised via a transformation induced only by targeting formant shifts. These transformations were obtained using a very simple VAE, suggesting that latent variable methods may be worth pursuing as simple and reproducible tool for sociolinguistic analysis which are consistent with formant normalization, but overcome formant measurement difficulties, and transform the whole spectrogram in a principled way. The simple method we investigated here could be improved first by developing a model of the full spectrum rather than the LPC envelope, and second by exploring the extent to which latent space normalisation changes physical rather than sociolingiustic features, with a view to developing latent space transformations optimally targeted to purely physical shifts.

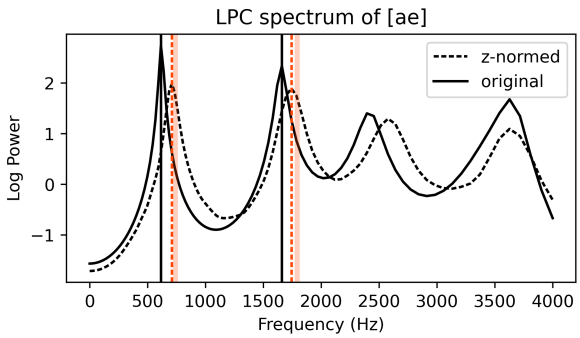

Figure 3: *Example spectrogram transformed by direct normalisation of latent features. Vertical back and red-dashed lines show original and shifted formants. Light red vertical lines show standard Lobanov normalised formant values.*

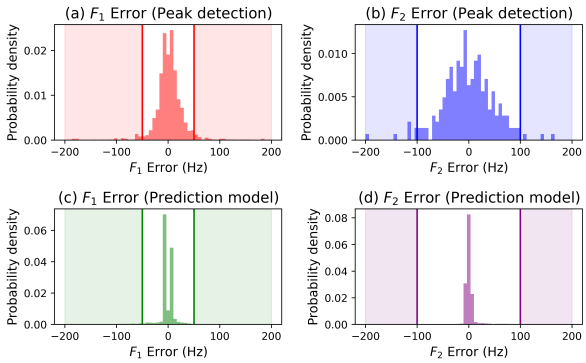

Figure 4: *Distribution of errors in formants of shifted spectrogram with respect to targets computed via (a,b) peak detection (c,d) prediction model applied to shifted latent features. Shaded regions indicate ranges of standard Lobanov shifts considered.*

5. References

[1] W. Labov, *Principles of Linguistic Change*. Oxford: Blackwell, 1994.

[2] J. Chambers and P. Trudgill, *Dialectology*. Cambridge: Cambridge University Press, 1998.

[3] D. Fogerty and D. Kewley-Port, "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 126, 2009.

[4] F. Chen, L. Wong, and E. Wong, "Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, 2013.

[5] S. Disner, "Vowel quality: The relation between universal and language-specific factors," *UCLA Working Papers in Phonetics*, vol. 58, 1983.

[6] H. Orton and E. Dieth, *Survey of English dialects*. Leeds: E. J. Arnold, 1962.

[7] M. Huckvale, "Accdist: a metric for comparing speakers' accents," in *Proc. INTERSPEECH 2004*, Jeju, Korea, Oct. 2004, pp. 100–104.

[8] O. Garcia, N. Flores, and M. Spotti, *The Oxford Handbook of Language and Society*. Oxford: OUP, 1994.

[9] K. Johnson and M. Sjerps, "Speaker normalization in speech perception," *UC Berkeley PhonLab Annual Report*, vol. 14, 2018.

[10] P. Roach, *English Phonetics and Phonology*. Cambridge University Press, Cambridge, 2009.

[11] G. E. Peterson and H. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, p. 175, 1952.

[12] J. Benesty, M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin: Springer, 2008.

[13] I. R. Titze *et al.*, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015.

[14] C. H. Shadle, H. Nam, and D. H. Whalen, "Comparing measurement errors for formants in synthetic and natural vowels," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 712–727, 2016.

[15] J. Lee, S. Shaiman, and G. Weismer, "Relationship between tongue positions and formant frequencies in female speakers," *Journal of the Acoustical Society of America*, vol. 139, p. 426, 2016.

[16] http://www.internationalphoneticassociation.org/content/ipa-chart, 2015.

[17] D. Jones, *An English Pronouncing Dictionary*. London: Dent, 1917.

[18] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, p. 346, 1996.

[19] N. Flynn and P. Foulkes, "Comparing vowel formant normalization methods," in *International Congress of Phonetic Sciences*, 2011.

[20] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[21] C. G. Clopper, D. B. Pisoni, and K. de Jong, "Acoustic characteristics of the vowel systems of six regional varieties of american english," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1661–1676, 2005.

[22] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *The Journal of the Acoustical Society of America*, vol. 67, pp. 971–994, 1980.

[23] J. Allen, M. S. Hunnicutt, and D. H. Klatt, *From Text to Speech*. Cambridge: Cambridge University Press, 1987, pp. 108–122.

[24] M. Furukawa, N. Wakatsuki, T. Ebihara, Y. Maeda, and K. Mizutani, "Vocal tract shape estimation for speech imitation by motorized vocal tract models," in *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*, 2022, pp. 505–508.

[25] M. Van Soom and B. de Boer, "A new approach to the formant measuring problem," *Proceedings*, vol. 33, no. 1, 2019.

[26] W. Styler, "On the acoustical features of vowel nasality in english and french," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, 2017.

[27] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, pp. 307–392, 2019.

[28] B. M and B. J., "Modeling and transforming speech using variational autoencoders," *Proc. Interspeech 2016*, pp. 1770–1774, 2016.

[29] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning Latent Representations for Speech Generation and Transformation," in *Proc. Interspeech 2017*, 2017, pp. 1273–1277.

[30] S. Kumar, S. Rath, and A. Pandey, "Speaker normalization using joint variational autoencoder," *Proc. Interspeech 2021*, pp. 1289–1293, 2021.

[31] S. Tan and K. C. Sim, "Learning utterance-level normalisation using variational autoencoders for robust automatic speech recognition," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 43–49.

[32] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, and R. Séguier, "Learning and controlling the source-filter representation of speech with a variational autoencoder," *Speech Communication*, 2023.

[33] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[34] 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*, vol. 8, pp. 185–190, 1937.

[35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cdrom. nist speech disc 1-1.1," NASA STI/Recon Technical Report No. 93, Tech. Rep., 1993.

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[38] https://github.com/james-burridge/latent_space_norm.