INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Analyzing intra-speaker and inter-speaker vocal tract impedance characteristics in a low-dimensional feature space using t-SNE

Balamurali B T, Jer-Ming Chen

Audio Research Group, Singapore University of Technology & Design, Singapore
balamurali_bt@sutd.edu.sg, jerming_chen@sutd.edu.sg

Abstract

In an earlier study [1], we have successfully classified a vowel-gesture parameter, gamma $\gamma(f)$ (relative vocal tract impedance spectrum measured using broadband signal excitation applied at the speaker's mouth during vowel phonation), via ensemble classification yielding accuracy exceeding 80% for six nominal regions of the vowel plane. In this follow-up investigation, we analyze gamma using t-SNE, a dimension reduction technique to allow visualizing gamma in low dimensional space, at two levels: inter-speaker and intra-speaker. Examining the same gamma dataset from [1], t-SNE yielded good spatial clustering in identifying the 6 different speakers with an accuracy exceeding 90%, attributable to the inter-speaker variation. Next, we further evaluated gamma of measurements only from a particular speaker in the lower dimension, which indicates intra-speaker distribution which may be associated with different measurement sessions. Using gamma may be seen as a meaningful parameter deserving further study, because it is inherently a function of the calibration load – unique for every speaker and measurement session. Because the calibration is made with the subject's mouth closed, so the measurement field during calibration is loaded solely by the impedance of the radiation field as seen at the subject's lips and baffled by the subject's face (geometrical information).

Index Terms: Vocal tract Impedance, Relative impedance spectrum, t-SNE, Inter (and intra) speaker variance.

1. Introduction

This study follows from one reported at Interspeech 2018 [1], where a relative vocal tract impedance spectrum parameter $\gamma(f)$ was used to successfully predict the vowel class label for $\gamma(f)$ associated with different nominal regions of the vowel plane.

The target parameter $\gamma(f)$ is derived by applying an external broadband excitation to the speaker's lips during phonation of 17 English vowels [2-4] following a calibration measurement that accounts for the acoustic radiation load associated with the speaker's facial geometry taken with lips closed; this technique was initially developed to estimate vocal tract resonance during singing and speaking [5-8]. A distinction here from [5-8] is the measurement hardware now used has been reduced and simplified, such that the vocal tract measurements can be easily deployed 'in the field' to facilitate a more 'ecological/natural' tracking of phonatory gestures.

Based on the success of [1], we started examining other possible methods of analyzing $\gamma(f)$'s feature space for various speakers and vowels. Accordingly, we report here an approach using t-Distributed Stochastic Neighbor Embedding (t-SNE) [9]**,** a nonlinear dimensionality reduction technique, which allows high-dimensional data to be visualized in a low-dimensional space. This allows us to identify features in $\gamma(f)$ which may not be immediately obvious in the earlier approach [1]. Perhaps in addition to classifying for vowel used, speaker identification and session separation are also possible.

The remainder of this paper is organized as follows: Section 2 describes the data collection procedure followed in this investigation. This include a brief overview of incorporated hardware, speaker participated and data collection protocol. A short description t-SNE can be found in Section 3. Details of gamma analysis at various levels is presented in Section 4. Finally, conclusions from this investigation is presented at Section 5.

2. Data Collection

2.1. Hardware

The hardware used in this study ("ACUZ-lite") is a compact handheld version of that reported earlier in [2-4], now with enhanced acoustic coupling and new electronics to allow greater signal:noise detection at the low frequency limit (~10dB boost @200 Hz) and reduced sampling time of 0.75 seconds. Description of the hardware, technique and acoustics is detailed here [10].

The broadband excitation signal consists of harmonics spaced 5.383 Hz (44100 Hz/(2^13)) apart, between 200 and 4000 Hz, summed, with optimized phases to improve the signal to noise [11], and delivered to a portable Nakamichi "mini cube" speaker. This source of acoustic flow is placed at the speaker's lips, with the lips resting against the speaker grill. Also located on the grill is a small electret microphone (Optimus 33-3013) which records both the sound of the speaker's voice along with the excitation signal interacting with the subject's vocal tract and the radiation field. In this way, each $\gamma(f)$ measurement collects both the relative impedance and relative phase information containing the first, second and third speech resonances associated with each vowel gesture.

At the start of the session with each subject, an initial measurement calibration is made with the subject's mouth closed: the measurement field here is loaded purely by the impedance of the effective radiation field as seen at the subject's lips, baffled by the subject's face, and the signal adjusted such that the resulting measured pressure signal is independent of frequency. Once calibrated, subsequent measurements of phonation gesture are then made with the subject vocalizing and positioned naturally (i.e., the lips open) to yield the relative vocal tract impedance spectrum parameter $\gamma(f)$, a ratio of the vocal tract impedance operating in parallel with the radiation field, with respect to the earlier calibration measurement, in response to the same acoustic flow.

http://dx.doi.org/10.21437/Interspeech.2019-1492

2.2. Subjects

6 speakers (4 men, 2 women) were recruited as part of this study [1]: four are native English Speakers (two speaking Australian English and two speaking Singaporean English), while two are non-native English Speakers. Ethnically, four are of European extraction and two have East Asian origin.

2.3. Data Collection Protocol

As reported earlier in [1], each subject was asked to phonate words of the form: h-*V*-d, where *V* is the target vowel between the consonants "h" and "d". 17 target vowels were used: 13 English vowels and 4 rhotacized/retroflex vowels. The target words are as follows: heed, hid, head, haired, haiRed, who'd, herd, heRd, hud, hard, haRd, had, hood, hoe'd, hoard, hoaRd, and hod. (Here, the uppercase "R" indicates rhotacized/retroflex version of the vowel sound is used.)

A gamma measurement was made for every word while holding the target vowel for about 2-3 seconds. With 4-5 'takes' of each target vowel, 17 target vowels: ~65 vowel gestures were collected from every subject. Further, for one of the women speakers (speaking Australian English), two more non-contemporaneous sessions of gamma measurements were made to study the intra-speaker variation.

3. t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) [9] is a well-known technique suited for the visualization of high-dimensional dataset. In t-SNE, a non-linear dimensionality reduction is achieved by converting similarities between data points to joint probabilities and minimizing the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. t-SNE can capture much of the local structure of the high-dimensional data very well. Additionally, it can also reveal the presence of clusters at several scales (i.e., global structure in the high dimensional data). Such local and global structures are revealed by putting emphasis on modelling dissimilar data points by means of large pairwise distances, and modelling similar data points by means of small pairwise distances. Researchers have previously applied t-SNE to speaker/ speech recognition arenas to create compelling low dimensional mappings of high dimensional feature space [12-14].

One of the key hyper parameters that tunes the balance between local and global structures present in the input data is perplexity. This parameter guesses number of close neighbors each point can have. The typical value for perplexity is chosen between 5 and 50 and in our investigation, we chose it to be 30 [15]. Joint distribution of input data and the gradient for optimization can be computed either approximately (i.e., using Barneshut approximation) or exactly. In this investigation, we used the latter procedure. Finally, when calculating distance between observations we used Minkowski distance.

4. Gamma Analysis

4.1. Study 1: Inter-speaker Gamma

Figure 1 shows the resulting relative impedance spectrum *γ*(*f*) (amplitude) for all the six speakers for the target vowel sound [ə] while phonating the word 'Herd'. For visual clarity, the extracted spectrums here are post-processed using Savitzky-Golay FIR smoothing filter [16] to remove artefacts associated with fundamental frequency information in the speech signal. However, while analyzing gamma using t-SNE we have used the original spectrum as such with no post processing.

To date, the gamma spectrum is used only to identify vocal tract resonances [2-8]. As expected, the gamma spectra shown here in Figure 1 indeed clearly indicates vocal tract resonances present (associated with a sharp negative slope [10]) at roughly ~500 Hz, 1500 Hz and 2500 Hz for the 6 speakers (which is expected for the target vowel [ə]). All inter-speaker variation in the gamma spectra can be clearly observed in Figure 1, and this arises due to the rather different facial and vocal tract physiology of each speaker. In order to visualize these differences in gamma associated with various phonemes phonated by six speakers in this investigation, the original gamma has been transformed to a lower dimension using t-SNE. Gamma here has a resolution of 5.383 Hz (Section 2.1) and to analyze a region between 200 and 4000 Hz, the resulting gamma will have 700 points. This high dimension gamma corresponding to all 6 speakers are then transformed to three dimensions, as shown in Figure 2.

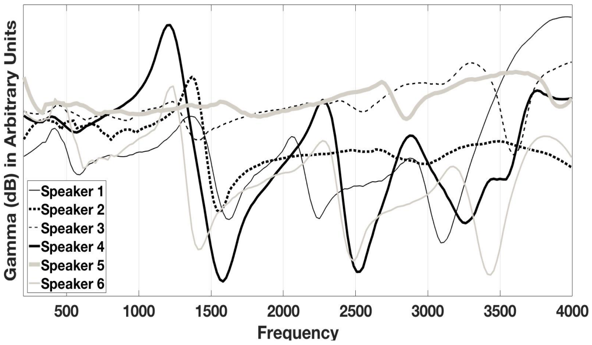


Figure 1: Relative impedance spectrum *γ*(*f*) (amplitude) for the target vowel [ə], measured from six speakers while phonating the word 'Herd', showing the wide variety of gamma spectrum structure observed across speakers while depicting somewhat consistent vocal tract resonances.

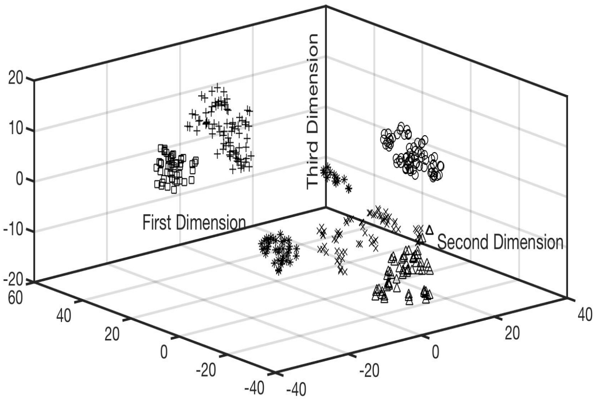


Figure 2: TSNE distribution of Gamma: ∗ – Speaker1, o – Speaker2, + – Speaker 3, × – Speaker 4, □ – Speaker 5, Δ – Speaker 6.

The relative impedance spectrum in this low dimensional embedding has resulted in a number of well-separated clusters. When examined further, the points corresponding to each cluster are identified to belong to a particular speaker. Most of the clusters are found to be tight (i.e., low variance within group) except one (that of speaker 1) where the points are dispersed somewhat loosely, but nevertheless still separated from the other clusters.

In an unsupervised manner, we have found out useful structure in this transformed gamma. Transformed gamma in lower dimension $\gamma^{(i)(j)}$; i = 1, . . . , n; j = 1,2,3 where $\gamma^{(i)} \in R^d$, n is the number of recordings; had resulted in set of clusters $C_1, \ldots, C_k$. In this investigation, it was found that integer K is equal to the number of speakers. However, t-SNE transformation is non-linear and often adapts (performing different transformations on different regions) to the given underlying data. As a result, the obtained cluster (See Figure 2) might not be well separated with the addition or deletion of data point thereby questioning the usage of such clusters to identify the speaker identity. In order to test this latter hypothesis, we conducted an experiment for which 80% of the available recordings across all speakers was used for obtaining the lower dimensional embedding (This 80% recordings will be referred as known recordings in this investigation). The remaining 20% of the recordings (~13 tested recordings per speaker) were tested against the low dimensional embedding to find whether they can be assigned to their original cluster. We refer this part of the study as "cluster assignment".

4.1.1. *Cluster Assignment - Methodology*

An unknown data point is assigned to a particular cluster as follows. First, a low dimensional embedding is created using 80% of the available recordings. We notice that gamma in this low dimensional embedding results in clusters.

The centroid of each cluster (designated as z_k) was then determined. This was chosen to be that point for which the sum squared distances from each point in a cluster to this chosen point was minimum. Approximately, this centroid will be closer to the center of ellipses shown in Figure 3. This later assumption would be true for every ellipse except the one drawn using dotted line. For this case, the centroid for the cluster was found to be closer towards the edge of the bigger dotted ellipse.

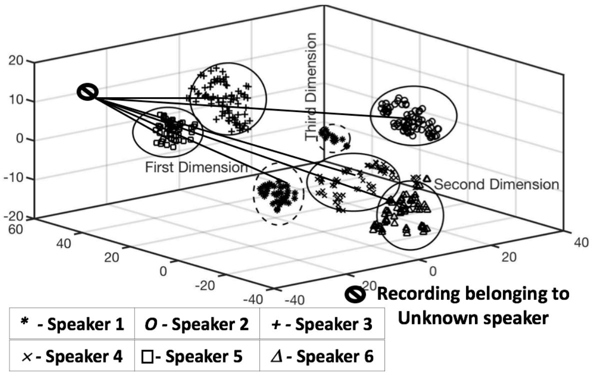


Figure 3: Speaker Space identified in a low-dimensional space derived from Gamma

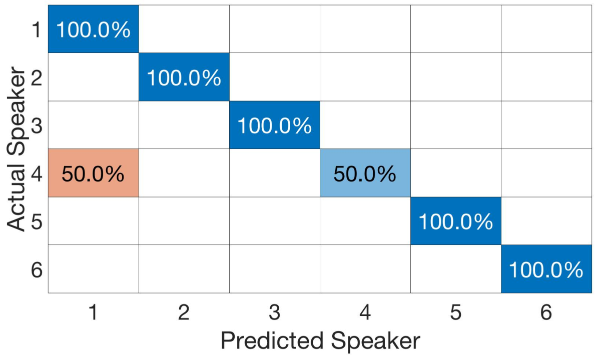
In order to find speaker identity of an unknown recording $\gamma_u(i)$; i = 1, firstly, the whole known recordings and the unknown recording is transformed using t-SNE to a lower dimension. Once the location of the unknown recording in the lower dimensional space is identified, the distance of its location to centroid of all the other recording (see Figure 3) is determined. The unknown recording was assigned to that particular speaker whose centroid happens to be the closest neighbor to the unknown recordings. Euclidian distance ($Euclidean\ dist\ ({\gamma_u}^i, z_k) = ({\gamma_u}^i - z_k)({\gamma_u}^i - z_k)'$) is used in this investigation to find this measure of closeness.

A new t-SNE transformation is required for every prediction, which can be highly time consuming when the data set contains large number of recording samples and this is indeed one of drawbacks with this procedure.

4.1.2. *Cluster Assignment - Result*

The accuracy of aforementioned cluster assignment is determined by comparing assigned labels with the actual labels. A confusion matrix [1] was created for analyzing this accuracy and this is shown in Figure 4. It is clear that speakers are often assigned to correct clusters except for one of the cases for which Speaker 4 is wrongly assigned as Speaker 1 half of the time. This is because Speakers 1 and 4 are located closely in the low dimensional embedding space. Further, the intra-speaker variance for Speaker 1 is higher compared to other clusters associated with other speakers (see Figure 3). As a result, when measure of closeness is estimated, some gamma from Speaker 4 are wrongly assigned as Speaker 1. It may be reasonable to surmise that such misclassifications rise with increased number of speakers in a given database. Nevertheless, this approach may still provide some degree of discrimination to facilitate meaningful inter-speaker identification.

Such inter-speaker discrimination may arise because the measurement is made at the subject's mouth, so the resulting gamma will contain acoustic information associated with geometrical features pertaining to the subject's physiology.



	1	2	3	4	5	6
1	100.0%					
2		100.0%				
3			100.0%			
4	50.0%			50.0%		
5					100.0%	
6						100.0%

Figure 4: Confusion matrix showing accuracy of Cluster Assignment (*cf*. Section 4.1).

4.2. Study 2: Intra-speaker Gamma

Here, we look at gamma corresponding to multiple sessions belonging to a particular speaker (Speaker 4). The motivation here is to see if, for a single speaker, there are variations associated with gamma collected across different measurement sessions. Figure 5 shows the resulting relative impedance spectrum $\gamma(f)$ (amplitude) for this particular speaker for three different non-contemporaneous sessions for the target vowel sound [ə] while phonating the word 'Herd'. Settings for Figure 5 were kept consistent as that for Figure 1.

Here, the intra-speaker variation present in impedance spectrum in Figure 5 can be observed, though still certainly more consistent when compared to the inter-speaker variation across different speakers (*cf.* Figure 1). Finally, the low dimensional embedding of gamma was created (*cf.* Section 4.1) to visualize the nuanced features (if any) in intra-speaker gamma distribution and is shown in Figure 6.

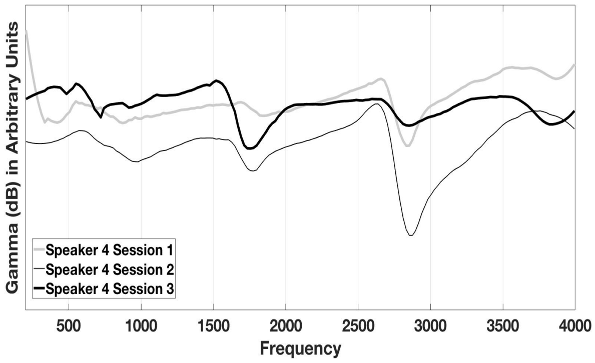


Figure 5: Relative impedance spectrum $|\gamma(f)|$ corresponding to a particular speaker corresponding to three non-contemporaneous session for the target vowel [ə] measured while phonating the word 'Herd'.

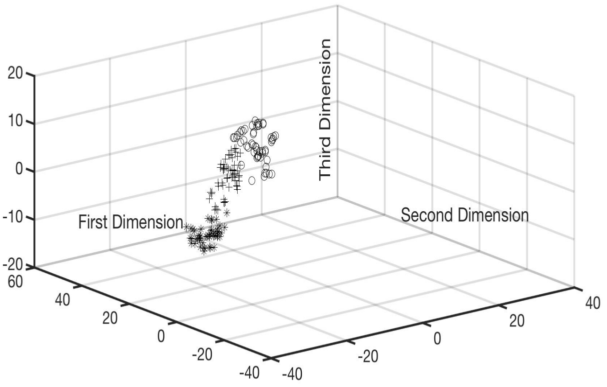


Figure 6: TSNE distribution of Gamma for Speaker 4: $\ast$ – Session1, o – Session2, + – Session3

Figure 6 shows that Gamma in the low dimensional feature space for a particular speaker is distributed quite tightly as a cluster and this is of no surprise (*cf.* Section 4.1). However, upon looking carefully within the cluster, one can further see non-overlapping regions (with some exceptional data points) associated clearly with each measurement session. Because the measurement calibration for gamma is made with the subject's mouth closed, so the measurement field during calibration is loaded solely by the impedance of the radiation field as seen at the subject's lips and baffled by the subject's face, and so may be attributed to natural variations associated with the calibration step for each session.

5. Conclusions

We have shown that t-SNE can be meaningfully implemented on a novel speech parameter, gamma (relative vocal tract impedance spectrum) to derive a low dimensional feature space, and offer analysis not previously studied in [1] such as inter- and intra-speaker variations. Such inter-speaker discrimination may arise because gamma contains acoustic information associated with geometrical features pertaining to the subject's face and vocal physiology. Further, for gamma data collected for the same speaker across sessions, meaningful intra-speaker regions may be identified, which are likely due to subtle variations associated at the calibration step during each session.

6. Acknowledgements

Part of this work was funded by the Australian Research Council (ARC) Linkage Project Grant (LP110100570) undertaken at the University of New South Wales with Professors Joe Wolfe and John Smith. We also thank our volunteer subjects.

7. References

[1] B. Balamurali and J.-M. Chen, "Automated Classification of Vowel-Gesture Parameters using External Broadband Excitation," in *Interspeech*, Hyderabad, India, 2018.

[2] J. Epps, J. Smith, and J. Wolfe, "A novel instrument to measure acoustic resonances of the vocal tract during phonation," *Measurement Science and Technology,* vol. 8, no. 10, p. 1112, 1997.

[3] A. Dowd, J. Smith, and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time," *Language and Speech,* vol. 41, no. 1, pp. 1-20, 1998.

[4] Y. Swerdlin, J. Smith, and J. Wolfe, "The effect of whisper and creak vocal mechanisms on vocal tract resonances," *The Journal of the Acoustical Society of America,* vol. 127, no. 4, pp. 2590-2598, 2010.

[5] M. Garnier, N. Henrich, J. Smith, and J. Wolfe, "Vocal tract adjustments in the high soprano range," *The Journal of the Acoustical Society of America,* vol. 127, no. 6, pp. 3771-3780, 2010.

[6] E. Joliveau, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: The soprano voice," *The Journal of the Acoustical Society of America,* vol. 116, no. 4, pp. 2434-2439, 2004.

[7] E. Joliveau, J. Smith, and J. Wolfe, "Acoustics: tuning of vocal tract resonance by sopranos," *Nature,* vol. 427, no. 6970, p. 116, 2004.

[8] T. Donaldson, D. Wang, J. Smith, and J. Wolfe, "Vocal tract resonances: a preliminary study of sex differences for young Australians," *Acoustics Australia,* vol. 31, no. 3, pp. 95-98, 2003.

[9] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, no. Nov, pp. 2579-2605, 2008.

[10] Thilakan Jithin, Balamurali B T and Jer-Ming Chen, "ACUZ-Lite: Ultra-Portable Real-Time Estimation of Vocal Tract Resonance," in *WESPAC* New Delhi, 2018.

[11] J. R. Smith, "Phasing of harmonic components to optimize measured signal-to-noise ratios of transfer functions," *Measurement Science and Technology,* vol. 6, no. 9, p. 1343, 1995.

[12] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *The Journal of Machine Learning Research,* vol. 15, no. 1, pp. 3221-3245, 2014.

[13] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, 2016, pp. 1-6: IEEE.

[14] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication,* vol. 81, pp. 54-71, 2016.

[15] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill,* vol. 1, no. 10, p. e2, 2016.

[16] H. H. Madden, "Comments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data," *Analytical chemistry,* vol. 50, no. 9, pp. 1383-1386, 1978.