An Unsupervised Speaker Identification Approach: A Breakthrough 3D Visualization of High Dimensional Features

Omar Elnaggar, Roselina Arelhi

***Abstract*—**An approach for unsupervised speaker identification based on text dependent speaker-specific phoneme model is proposed. The highlight of the approach is the use of parametric t-distributed stochastic neighbour embedding (t-SNE), to map high dimensional extracted features to a lower dimension such that intrinsic features are retained. This is the first time that parametric t-SNE was successfully implemented to visualize high dimensional voice features in a 3D space. The visualisation facilitates tuning of the parameters of the speaker models based on the topology of the speakers' low dimensional features. The experimental results for a speaker identification set up using the CSLU ISOLET Spoken Letter Database shows a 3D plot of clusters with good class sparsity, measured using the Kullback-Liebler divergence metric. As no other similar work has been reported, a fair model-to-model comparison of performance is not available. However, the rate achieved is comparable to other works using parametric t-SNE for automatic classification of data in other domains.

***Keywords*—**deep neural network, gap statistic, gaussian mixture model, parametric t-distributed stochastic neighbor embedding, speaker identification.

I. Introduction

SPEAKER recognition is a technique to automatically identify who is speaking. Automatic speaker recognition systems have existed since the 1960s with early works by [1]–[3]. It can be classified into two general tasks; *speaker identification* and *speaker verification*. In speaker identification [4]–[6], utterance from an unknown speaker is compared with utterances of known speakers in a closed set. The unknown speaker is identified as the most likely speaker with the highest similarity score. For speaker verification [7][8], the system only checks for similarity with utterances associated with the claimed identity in the closed set. The similarity score is compared with a threshold to make an accept or reject decision. The threshold is set to give a good trade-off between accepting impostors and falsely rejecting valid speaker.

Speaker recognition techniques can be divided into text dependent and text independent techniques. In text-dependent systems, the texts are fixed, and it is assumed that the speaker would utter the same texts as for training. These constraints are quite reasonable and can improve the accuracy of the system. Many applications such as biometric systems to allow authorized access are based on scenarios with cooperative users speaking fixed texts. However, there are also applications where it is not possible to enforce such constraints. One example is in a forensic identification system, where the identity of the person in question is identified behind the scene. A system with no constraints on the spoken texts is known as text-independent systems. Since the speaker can utter any words or phrases during recognition, it is more difficult to fool the system as it is harder to mimic unknown texts. Hence, text-independent systems are more challenging [7].

This paper proposes an approach for speaker identification, which could be used for many speaker recognition tasks based on short utterances. Central to the approach is the use of the dimensionality reduction technique, *parametric t-distributed stochastic neighbour embedding* (pt-SNE) which, for the first time, successfully map high dimensional extracted voice features to a low dimensional space with good class separability. The advantage of using the technique is that it allows embedding out-of-sample data without the need to re-train the map. Also, utilising a *deep neural network* (DNN) allows for quick embedding of high-dimensional features into its pre-trained low-dimensional map, computational efficiency and improved decision-making duration. The paper is organized as follows. In Section II, the proposed approach and related work which gives the theoretical background for our work is presented. Section III presents the voice corpus used and the experimental set up. Section IV describes the optimization strategy for determining the optimal DNN architecture and observations made from the successive visual analysis. Sections V and VI discuss results for the speaker modelling and decision-making stages respectively while Section VII concludes the paper with recommendations for future work.

II. Proposed Approach and Related Work

Fig. 1 shows the proposed approach [9] which incorporates and integrates existing techniques; specifically, the pt-SNE.

A. Pre-processing

A speaker identification system begins with a training stage where users provide utterances for training. These are passed through a pre-processing stage to provide features suitable for speaker modelling. The utterances are divided into frames of 25ms length using Hamming window at 10ms frame rate. A

Omar Elnaggar is with the Electrical and Electronic Engineering Department, University of Nottingham, Jalan Broga, Semenyih 43500, Selangor, Malaysia (phone: +60 3 8924 8120; e-mail: kecy5oms@nottingham.edu.my).

Roselina Arelhi is with the Electrical Engineering Department, University of Nottingham, Jalan Broga, Semenyih 43500, Selangor, Malaysia (phone: +60 3 8924 8120; e-mail: roselina.arelhi@nottingham.edu.my).

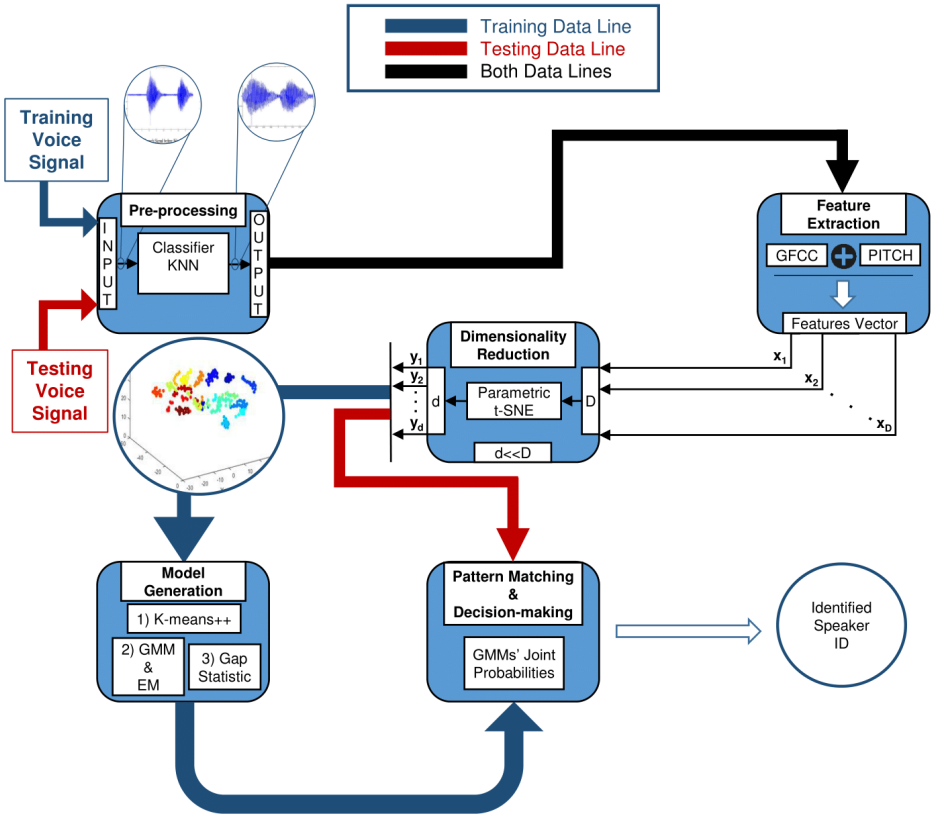

Fig. 1 Proposed Unsupervised Speaker Identification Approach

voice activity detection proposed by [10] is then performed to remove silence segments from the samples to ensure that they do not contribute to clusters in the visualization stage.

B. Feature Extraction

The next stage extracts salient features conveying speakers' information. We used gammatone frequency cepstral coefficient (GFCC) applied to the frames to provide 66 coefficients. Even though our voice data (presented in Section III) do not contain noise, we used GFCC for the following reasons. GFCC minimizes sensitivity to noise - an important feature for a real scenario where speakers would usually utter texts in a noisy environment. Furthermore, the Gaussian Mixture Models (GMM) [11] which we will employ for speaker modelling in a later stage usually do not perform well under noisy conditions [12][13]. Therefore, the use of GFCC in an earlier stage would reduce performance degradation by GMM. We also found that using GFCC instead of the more common MFCC, provided less class-overlapping on 3D visualization graphs obtained using the t-SNE dimensionality reduction technique. One reason for the better performance is that GFCC has an Equivalent Rectangular Bandwidth (ERB) scale of a finer resolution compared to the Mel scale over the low-frequency range, where most of the speech energy resides [13]. The features provided by GFCC are appended with pitch, a speaker's voice attribute, to give a total of 67 features. Pitch is a gender-dependent feature that will provide a more precise characterization of speakers to increase overall identification rates [14] [15].

C. Dimensionality Reduction

Visualization of the 67 features on a low dimensional map would facilitate identification of speakers. This brings the need for a dimensionality reduction technique to map the high dimensional features to a lower dimension such that the intrinsic characteristics of speakers are retained. A dimensionality reduction technique reduces the high dimensions of an $n \times D$ dataset $\mathbf{X}$ into a new $n \times d$ low dimensional dataset $\mathbf{Y}$, where n is the number of datavectors $\mathbf{x_i}$ and $\mathbf{y_i}$ $(i \in \{1,2,\dots,n\})$, D is the original dimensionality of $\mathbf{X}$ and d is the *intrinsic dimensionality* (usually $d \ll D$). The term, *intrinsic dimensionality*, indicates that datavectors $\mathbf{x_i}$ lie on or near a d-dimensional manifold embedded in the D-dimensional space [16].

Some early dimensionality reduction techniques are *principal component analysis* (PCA) [17], *linear discriminant analysis* (LDA) [18][19], *multidimensional scaling* [20], *Isomap* [21], *local linear embedding* (LLE) [22] and *stochastic neighbour embedding* (*SNE*) [23]. In 2007, the prize-winning t-SNE, adapted from SNE, was introduced. t-SNE's key feature which captures local structure and preserve global distribution of features has made it one of the best techniques for visualizing high dimensional data as 2D or 3D scatterplots. It has been successfully applied to visualize data in many domains such as handwritten digits and images [24], text documents [25], genetic diseases [26] and animal mapping behaviour [27].

The primary feature of t-SNE is capturing of local structures, while retaining global data distribution. Its algorithm begins by

converting the Euclidean distances between pairs of datapoints $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ ($i \neq j$) in the high dimensional space (hyperspace) to the joint probability distribution $p_{j|i}$, the conditional probability that $\boldsymbol{x_j}$ would pick $\boldsymbol{x_i}$ as its neighbor.

$$p_{j|i} = \frac{\exp\left(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2}{2\boldsymbol{\sigma_i^2}}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_k}\|^2}{2\boldsymbol{\sigma_i^2}}\right)}, p_{i|i} = 0 \qquad (1)$$

where $\boldsymbol{\sigma_i}$ is the standard deviation of the Gaussian probability distribution centered about $\boldsymbol{x_i}$. Technically, dense regions have smaller values of $\boldsymbol{\sigma_i}$ compared to sparser regions. Any value of $\boldsymbol{\sigma_i}$ induces a probability distribution $\boldsymbol{P_i}$ over all the other datapoints. t-SNE carries out a binary search for its value based on a user-defined parameter called *perplexity* which can be interpreted as the average number of neighbor datapoints surrounding each $\boldsymbol{x_i}$. Values of the perplexity is typically varied between 5 and 50 till the optimum value is reached for the problem in hand [28].

$$\text{Perplexity}(\boldsymbol{P_i}) = 2^{H(\boldsymbol{P_i})} \qquad (2)$$

where $H(\boldsymbol{P_i})$ is the Shannon entropy of $\boldsymbol{P_i}$ measured in bits. Similar to p_{ij}, q_{ij} models the probability that map point $\boldsymbol{y}_j$, the low dimensional counterpart of $\boldsymbol{x_i}$, would take $\boldsymbol{y_i}$, as its neighbor using (3), following a Student t-distribution with one degree of freedom.

$$q_{ij} = \frac{\left(1+\|\boldsymbol{y_i} - \boldsymbol{y_j}\|^2\right)^{-1}}{\sum_{k \neq i}(1+\|\boldsymbol{y_i} - \boldsymbol{y_k}\|^2)^{-1}}, q_{i|i} = 0 \qquad (3)$$

The locations of $\boldsymbol{y_i}$ are determined by minimizing the cost function *C*, based on the Kullback-Leibler (*KL*) divergence [29][30][31] using a gradient descent method.

$$C = \sum_{\text{i}} KL(\boldsymbol{P_i}||\boldsymbol{Q_i}) = \sum_{\text{i}} \sum_{\text{j}} p_{ji} \log\left(\frac{p_{ji}}{q_{ji}}\right) \qquad (4)$$

where $\boldsymbol{P_i}$ represents the conditional probability distribution over all other datapoints given datapoint $\boldsymbol{x_i}$ and $\boldsymbol{Q_i}$ the equivalent for the low dimensional map.

A year after t-SNE was introduced, parametric t-SNE (pt-SNE) was proposed in 2009 [32]. The mapping from high dimensional space *X* to the low dimensional space *Y* is parametrized by means of a feed-forward deep neural network (DNN), which is trained in a way to preserve local structure of the data in *Y*. pt-SNE avoids being stuck in local minima by going through three consecutive stages for training its DNN: (i) training a stack of RBMs to minimize *contrastive divergence* [33][32], (ii) concatenating the stack of RBMs to construct the feedforward neural network and (iii) employing the standard backpropagation to minimize the cost function.

Like t-SNE, pt-SNE utilizes (1) to compute the Gaussian similarity conditional probabilities, p_{ij}, in the hyperspace. But since the latter technique is a parametric model, it computes Student-t distributed similarity probabilities, q_{ij}, in the low dimensional map as a function of high dimensional input datapoints $\boldsymbol{x_i}$ feedforward through the DNN being trained:

$$q_{ij} = \frac{\left(1 + \frac{\|\text{f}(\boldsymbol{x_i}|\boldsymbol{W}) - \text{f}(\boldsymbol{x_j}|\boldsymbol{W})\|^2}{v}\right)^{-\frac{\text{v}+1}{2}}}{\sum_{\text{k}\neq\text{l}}\left(1 + \frac{\|\text{f}(\boldsymbol{x_k}|\boldsymbol{W}) - \text{f}(\boldsymbol{x_l}|\boldsymbol{W})\|^2}{v}\right)^{-\frac{\text{v}+1}{2}}}, \quad q_{ii} = 0 \qquad (5)$$

where v is the number of degrees of freedom of Student-t distribution and $f(\boldsymbol{x_i}|\boldsymbol{W})$ refers to the DNN mapping of a high dimensional datapoint, $\boldsymbol{x_i}$, to a low dimensional map point, $\boldsymbol{y_i}$. There are multiple ways to decide about the variable v [32]; however, this paper considers the linear relationship governing the number of degrees of freedom such that $v = d - 1$. Next, the gradient of the cost function is partially differentiated with respect to the DNN's weights, $\boldsymbol{W}$, instead of map points in the case of t-SNE

$$\frac{\partial C}{\partial \boldsymbol{W}} = \frac{\partial C}{\partial f(\boldsymbol{x_i}|\boldsymbol{W})} \frac{\partial \text{f}(\boldsymbol{x_i}|\boldsymbol{W})}{\partial \boldsymbol{W}} \qquad (6)$$

where the differential term $\frac{\partial C}{\partial f(\boldsymbol{x_i}|\boldsymbol{W})}$ is given by

$$\frac{\partial \text{C}}{\partial \text{f}(\boldsymbol{x_i}|\boldsymbol{W})} = \frac{2(v+1)}{v} \sum_{\text{j}} (p_{ij} - q_{ij})\left[f(\boldsymbol{x_i}|\boldsymbol{W}) - f(\boldsymbol{x_j}|\boldsymbol{W})\right] \cdot \left(1 + \frac{\|f(\boldsymbol{x_i}|\boldsymbol{W}) - f(\boldsymbol{x_j}|\boldsymbol{W})\|^2}{v}\right)^{-\frac{v+1}{2}} \qquad (7)$$

and $\frac{\partial f(\boldsymbol{x_i}|\boldsymbol{W})}{\partial \boldsymbol{W}}$ can be computed using the standard backpropagation algorithm knowing the activation functions at each layer of the DNN. Consequently, each training epoch ends by updating the weights between each two successive layers.

E. Speaker Model Generation

The aim of speaker model generation is to obtain a unique identifier to the features of each speaker in the closed set of speakers. The techniques can be broadly categorized into template-based and stochastic-based techniques. Template-based techniques aim to minimize some squared distance or error measures. Examples are *Dynamic Time Warping* (DTW) [34] [35] and *Vector Quantization* (VQ) [36]. For stochastic-based techniques, each speaker is modelled as a probabilistic source with an unknown fixed probability function which are

estimated from the training data [37]. Matching is done by evaluating the likelihood of the test data with respect to the trained model. Examples of popular stochastic models are the *Hidden Markov Model* (HMM) and the *Gaussian Mixture Model* (GMM). Over the last decade, GMM has become established as the standard model for text-independent speaker recognition [11]. It is a parametric probability density function represented as a sum of Gaussian components densities. GMMs have three advantages (i) it is based on a well-understood statistical model (ii) it is computationally inexpensive and (iii) the models can be scaled and updated to add new speakers with relative ease [37] [38]. Our proposed approach uses GMM and includes using the *gap statistic* method proposed by Tibshirani et al [39] to determine the optimal number of clusters, hence Gaussian components, after evaluating the clustering performance against different numbers of clusters, and the *iterative expectation-maximization* (EM) algorithm proposed by [40] to perform the *maximum likelihood estimation* (MLE) [41] of the GMM's model parameters, which are initialized using K-means++ clustering algorithm [42] prior to applying the EM algorithm.

F. Decision-Making

Every out-of-sample datavector corresponding to testing extracted features is embedded into the pre-trained low-dimensional map, thus, compared against every mixture model based on its corresponding GMM's joint probability. Subsequently, each testing datavector is assigned to the speaker, whose GMM model yields the highest joint probability. Combining all the classification decisions of the overlapping testing frames, the current testing utterance is finally assigned to the most frequently identified speaker.

III. Experiments

t-SNE based techniques have shown more meaningful visualizations of high dimensional datasets compared to other techniques in terms of class separability and cluster conciseness. Some successful applications can be found in [43][44][45]. To-date, there has been no reported success on using pt-SNE for visualization of high dimensional voice features. Hence, our work investigated the feasibility of using pt-SNE as (i) it is a parametric technique, making it favored over other techniques for various applications, and (ii) it has many tunable parameters enhancing control over performance.

A. Data Corpus

The voice corpus used in this work is the ISOLET Spoken Letter Database version 1.3 [46], which consists of 75 male and 75 female speakers where each speaker holds two utterances of each English letter. Each utterance lasts for two seconds. After silence removal, the duration of each utterance varied, with a maximum of 1 second, depending on the length of silence in each recording. Speakers' ages range from 14 to 72 years old with an average of 35 years old.

B. Experimental Setup

The MATLAB R2017b was the software used for this experiment. The extracted features per frame of all speakers were normalized to values between 0 and 1 due to utilizing Bernoulli-distributed hidden nodes during RBM pre-training of the DNN. Each frame represents one high dimensional datapoint.

As t-SNE based techniques are unsupervised techniques, they do not take into consideration any class labelling of the input data. Since the aim is to classify purely speaker-dependent features, to rule out classification of speaker-independent features, the experiment was performed with one single letter, as the fixed text, for all speakers. Two utterances of the letter 'A' by each of the 16 speakers randomly selected from the database population, to produce visually color discriminant visualizations, were selected for training and testing phases respectively. The letter was selected as it is a vowel, which has been shown to have large inter-speaker variability and small intra-speaker variability, a desired feature for high performance of speaker recognition [47][48].

IV. *DNN* Optimization Strategy and Observations

pt-SNE has three parameters which can be tuned - the perplexity value, v and DNN architecture. For 3D visualizations of high dimensional datavectors, v would eventually be set to 2. Assuming initial value of 20 for perplexity, the DNN architecture is then determined. Our DNN architecture was motived by [32], and had an initial setting of $(D-100-100-100-d)$ where D is the number of sigmoid-activated neurons at the input layer, and d is the number of linear neurons which form the output layer. Hence for our data, D=67 and d=3. The DNN is trainable for 200 epochs. Fig. 2 shows the 3D visualization obtained for this initial architecture. A structure with moderate class separation and few overlapping is observed.

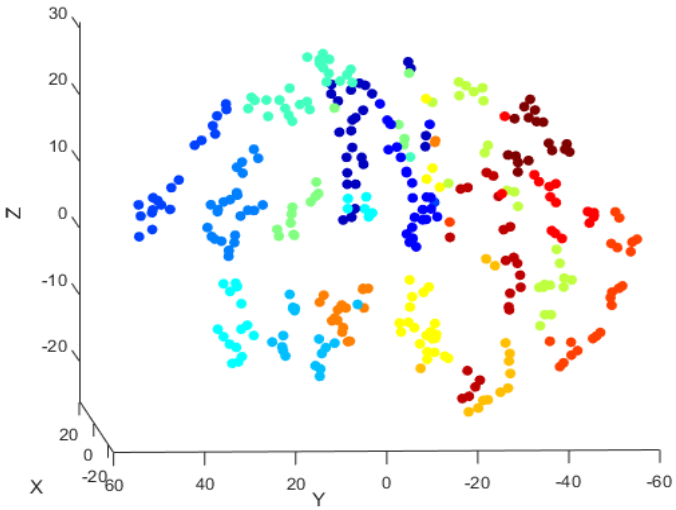

Fig. 2 A 3D view of low dimensional embeddings for letter 'A' only (DNN Architecture (67-100-100-100-3), KL Divergence: 34193e-5)

The DNN parameters are then optimized to improve class separation. The parameters are tuned according to the following priority order: (i) network's number of layers, (ii) number of

neurons employed at each layer, and (iii) maximum number of training epochs [49]. The quality of the low dimensional visualizations achieved were assessed using the *Kullback-Leibler* (KL) divergence obtained from five consecutive runs of pt-SNE for the same set of parameters and based on visual judgement of class separability and conciseness. The number of layers for DNN is a primary parameter determining its performance. Too few layers results in inadequate feature extraction from the network inputs and eventually lead to inaccurate network output(s). On the other hand, too many layers prevent generalization of the model as the DNN is over-trained to fit the training data only. Our experimental results for optimizing the DNN number of layers showed that a five-layer architecture succeeded in maximizing the global variance of low dimensional embeddings and was nominated as the optimum number of layers. To optimize the number of neurons, the three hidden layers were tested using combinations of 100, 200 and 300 neurons, giving 6 different possibilities. The five-run average KL divergence was evaluated for the 6 possible architectures to compare their performances. The DNN architecture (67-100-200-300-3) had a minimum average KL divergence value of 30214e-5. Preserving the pre-determined order of neuron-hungry layers, the number of neurons at each layer was varied up to 500 neurons[1] to estimate the optimum number of neurons for each layer. The five-run average KL divergence was again evaluated to compare performances. The DNN architecture (67-100-200-400-3) was chosen as it yields the minimum average KL divergence of 28654e-5. The corresponding low dimensional visualization of the selected architecture is shown in Fig. 3.

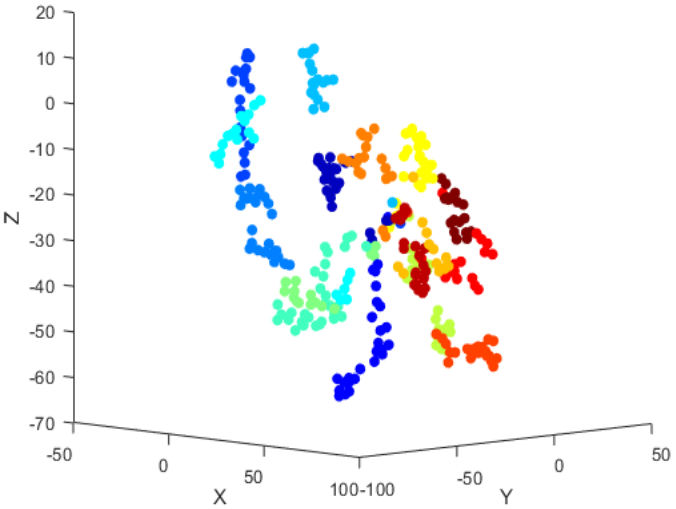

Fig. 3 DNN Architecture (67-100-200-400-3), KL Divergence: 28654e-5

The last parameter to be optimized is the maximum number of training epochs; one epoch is counted when all training datapoints are used once to update the DNN's weights. A DNN's learning curve constructed by the instantaneous KL divergences up to 1000 epochs was obtained and it was determined that 300 epochs are sufficient to achieve a compromise between training time and generalization of DNN performance for testing data.

Finally, pt-SNE utilizes the perplexity parameter to control the spread of neighborhood probability distribution centered about each datavector in the hyperspace. To visualize the effect of perplexity on the low-dimensional visualization, pt-SNE was run against several values of perplexity from 5 to 50. The best visualization was for perplexity value of 12 and is shown in Fig. 4. Both global and local evaluation criteria were met with clusters showing good separability. We noticed a very small region containing few points from each class, marked with a black circle. This could represent a short-duration feature, such as noise present during initial recordings, which is common to all speakers. As the overall quality of visualization was not affected, we have left removal of this feature as a future work to improve our pre-processing stage.

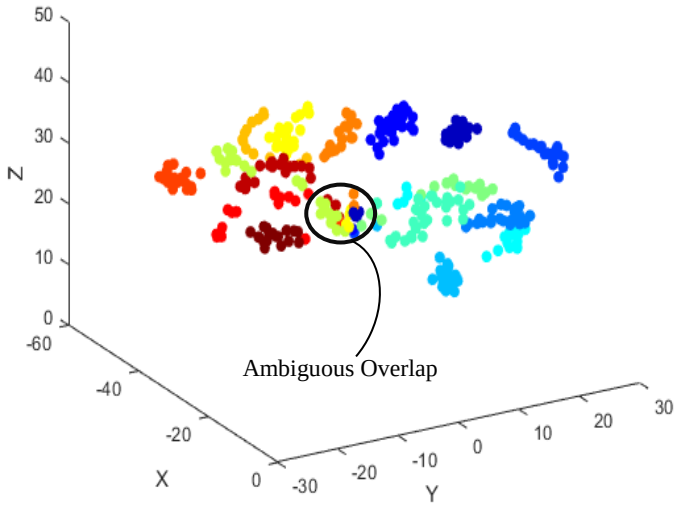

Fig. 4 Perplexity: 12; KL divergence 18509e-5

V. Speaker Modelling

Following our proposed approach, an unsupervised modelling of the 3D speaker-dependent features was conducted for each of the sixteen speakers (*Speaker 1* to *Speaker 16*). Fig. 5 shows samples of the visualizations for two speakers, indicating that the employed clustering approach is effective enough in terms of the clustering performance of GMM as well as the optimal number of clusters determined using the gap statistic. It can be noticed that an optimal number of clusters is assigned to each speaker based on his/her features' structure within the low-dimensional map, where each cluster's datapoints are having distinctive color.

[1] Poor visualization showed insufficient training of growing weighted connections, when employing more than 500 neurons

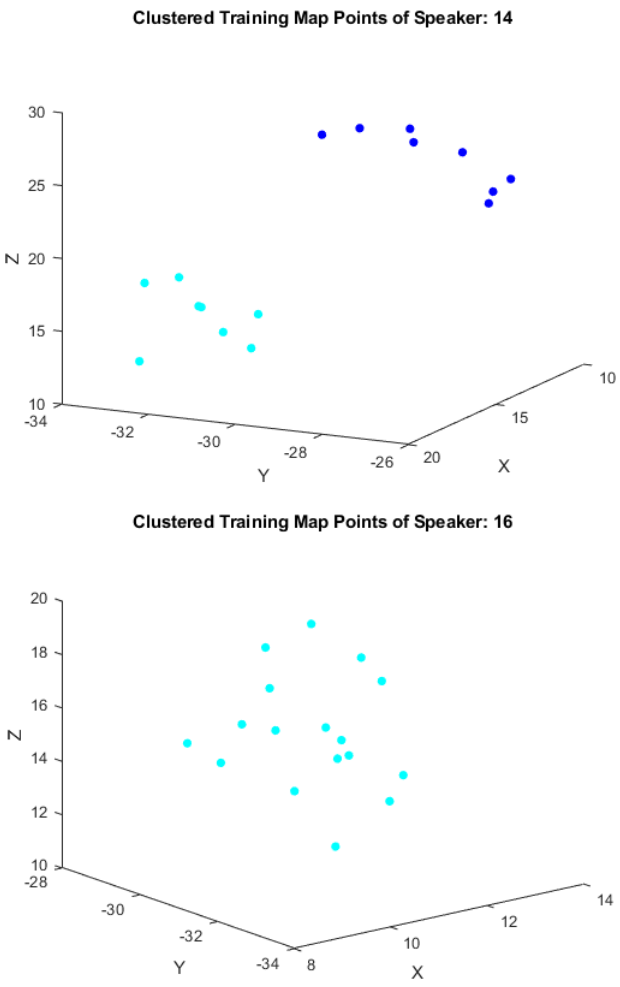

Fig. 5 Sample visualizations of unsupervised clustering of low-dimensional speaker-dependent features

VI. Decision Making

Given the system is completely unsupervised, the system's overall speaker identification accuracy slightly changes each time it is re-trained; however, it hit as high as 100% and 75% for training and testing datasets respectively. Comparing our classification accuracy with those reported in [50][51], it can be noticed that ours outperform many of the existing implementations of pt-SNE on other data domains. In addition, our utterances duration of less than 1 second is considered short and thus the accuracy rates achieved are acceptable with those reported in the literature for speaker recognition system based on short utterances [52].

VII. Conclusions and Future Work

As a summary, our contributions are: (i) a new approach for text dependent speaker identification which incorporates an unsupervised parametric t-SNE, and (ii) successful visualization of high dimensional features of voice signal in a 3D map. The advantage of using pt-SNE is that it allows embedding out-of-sample data without the need to re-train the map. Also, utilizing a DNN allows for effective low dimensional embeddings by feeding forward high dimensional features, in a matter of milliseconds, thus a more computationally efficient and fast decision-making approach. Although, for proof of concept we tested with a relatively small number of samples of voice signals, this is the first time that high dimensional features of a dynamic voice signal were successfully visualized in a 3D map using a parametric dimensionality reduction technique. As there is no similar work on voice signal, the best comparison that could be made was with implementations of pt-SNE on other data domains; the highest speaker identification accuracy of 75% we obtained for out-of-sample data was higher than the others. Further work can be performed with higher number of sample data though visual inspection would be challenging. Pt-SNE could be substituted with other variants of t-SNE such as kernel t-SNE [53] or fisher kernel t-SNE [53] which are also parametric techniques but they do not require inspection of visualization graphs by a human observer.

Acknowledgment

The authors wish to thank Wong Yee Wan for her advice on unsupervised learning and Chin Yong Sheng, Alimohamed Mussa and Wan Nazatul Alia Wan Zaki for their research support on the proposed approach diagram, GFCC and GMM.

References

[1] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J. Acoust. Soc. Am.*, vol. 35, no. 3, pp. 354–358, 1963.

[2] K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification Using an Adaptive system," *J. Acoust. Soc. Am.*, vol. 40, no. 5, pp. 966–978, 1966.

[3] B. S. Atal, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 72–82, 1971.

[4] P. Univaso, "Forensic Speaker Identification: A tutorial," *IEEE Lat. Am. Trans.*, vol. 15, no. 9, pp. 1754–1770, 2017.

[5] A. L. Higgins, L. G. Bahler, and J. E. Porter, "Voice identification using nearest-neighbor distance measure," *Acoust. Speech, Signal Process. 1993. ICASSP-93., 1993 IEEE Int. Conf.*, vol. 2, pp. 375–378, 1993.

[6] T. Kinnunen, E. Karpov, and P. Franti, "Real-Time Speaker Identification and Verification," *IEEE Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 277–288, 2006.

[7] F. E. E. Bimbot *et al.*, "A Tutorial on Text-Independent Speaker Verification," *EURASIP J. Appl. Signal Processing*, vol. 4, pp. 430–451, 2004.

[8] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-Independent Speaker Verification Using 3D

Convolutional Neural Networks," *arXiv:1705.09422*, pp. 1–6, 2017.

[9] O. Elnaggar, Y. S. Chin, A. A. Mussa, and W. N. A. Wan Zaki, "Multi-Modal Biometric Authentication System with Dimensionality Reduction," University of Nottingham, Malaysia, 2018.

[10] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis. A MATLAB Approach*. Academic Press, 2014.

[11] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[12] A. Drygajlo and M. El-Maliki, "Speaker Verification in Noisy Environments with Combined Spectral Subtraction and Missing Feature Theory," *IEEE Int. Conf. Acoust. Speech Signal Process. Acoust. Speech Signal Process.*, pp. 121–124, 1998.

[13] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7204–7208, 2013.

[14] S. P. Whiteside, "Sex-specific fundamental and formant frequency patterns in a cross-sectional study," *J. Acoust. Soc. Am.*, vol. 110, no. 1, pp. 464–478, 2001.

[15] L. M. Mazaira-Fernandez, A. Álvarez-Marquina, and P. Gómez-Vilda, "Improving Speaker Recognition by Biometric Voice Deconstruction," *Front. Bioeng. Biotechnol.*, vol. 3, pp. 1–32, 2015.

[16] L. van der Maaten, E. Postma, and J. van den Herik, "Dimensionality Reduction : A Comparative Review," 2009.

[17] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag. Ser. 6*, vol. 2, no. 11, pp. 559–572, 1901.

[18] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.

[19] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, 2004.

[20] T. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall, 1994.

[21] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "Isomap," *Science*, vol. 290, no. 5500, pp. 2319–23, 2000.

[22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding.," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[23] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," *Adv. Neural Inf. Process. Syst.*, pp. 833–840, 2002.

[24] W. M. Abdelmoula *et al.*, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data," *Proc Natl. Libr. Acad. Sci. USA*, vol. 113, no. 43, pp. 12244–12249, 2016.

[25] Y. Mao and K. Balasubramanian, "Dimensionality Reduction for Text using Domain Knowledge," *COLING '10 Proc. 23rd Int. Conf. Comput. Linguist. Posters*, no. August, pp. 801–809, 2010.

[26] W. Xu, X. Jiang, X. Hu, and G. Li, "Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization," *BMC Med. Genomics*, vol. 7, no. 2, p. S1, 2014.

[27] J. G. Todd, J. S. Kain, and B. L. De Bivort, "Systematic exploration of unsupervised methods for mapping behavior," *Phys. Biol.*, vol. 14, no. 1, 2017.

[28] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[29] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[30] S. Kullback, *Information Theory and Statistics*. 1959.

[31] S. Kullback, "Letter to the Editor: The Kullback–Leibler distance," *Am. Stat.*, vol. 41, no. 4, pp. 340–341, 1987.

[32] L. Van Der Maaten, "Learning a Parametric Embedding by Preserving Local Structure," *Proc. Twelfth Int. Conf. Artif. Intell. Stat. (AI-STATS), JMLR*, vol. 5, pp. 384–391, 2009.

[33] G. E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Comput.*, vol. 14, pp. 1771–2800, 2002.

[34] P. Senin, "Dynamic Time Warping Algorithm Review," *Science (80-.).*, vol. 2007, no. December, pp. 1–23, 2008.

[35] H. Sakoe and S. Chiba, "Dynamic Programming

Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

[36] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," *ICASSP '85. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 10, pp. 387–390, 1985.

[37] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process. A Rev. J.*, vol. 10, no. 1, pp. 19–41, 2000.

[38] K. Sarmah, "Comparison Studies of Speaker Modeling Techniques in Speaker Verification System," *Int J Sci. Res. Comput. Sci. Engienering*, vol. 5, no. 5, pp. 75–82, 2017.

[39] T. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Ser. B*, vol. 63, no. 2, pp. 411–423, 2001.

[40] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[41] N. E. Day, "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, vol. 56, pp. 463–474, 1969.

[42] D. Arthur and S. Vassilvitskii, "k-means ++ : The Advantages of Careful Seeding," *SODA '07 Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms*, vol. 8, pp. 1–11, 2007.

[43] S. Mounce, "Visualizing Smart Water Meter Dataset Clustering With Parametric T-distributed Stochastic Neighbour Embedding," *13th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.*, 2017.

[44] J. G. Todd, J. S. Kain, and B. L. de Bivort, "Systematic exploration of unsupervised methods for mapping behavior," *Phys. Biol.*, vol. 14, 2017.

[45] W. M. Abdelmoulaa *et al.*, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data," *Proc. Natl. Acad. Sci. United States Am.*, vol. 113, no. 43, 2016.

[46] R. Cole, Y. Muthusamy, and M. Fanty, "CSLU: ISOLET Spoken Letter Database Version 1.3 LDC2008S07," *Web Download. Philadelphia: Linguistic Data Consortium*, 2008. .

[47] J. P. Eatock and J. S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," *Proc. - ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process.*, no. June, pp. 134–136, 1994.

[48] M. R. Sambur, "Selection of Acoustic Features for Speaker Identification," *IEEE Trans. Acoust.*, vol. 23, no. 2, pp. 176–182, 1975.

[49] Satish Kumar, *Neural Networks - A Classroom Approach*. 2004.

[50] A. Gisbrecht, A. Schulz, and B. Hammer, "Parametric nonlinear dimensionality reduction using kernel t-SNE," *Neurocomputing*, vol. 147, pp. 71–82, 2015.

[51] L. Van Der Maaten, "Learning a Parametric Embedding by Preserving Local Structure," *JMLR Proc. vol. 5*, pp. 384–391, 2009.

[52] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker Identification and Clustering Using Convolutional Neural Networks," *Mach. Learn. Signal Process.*, vol. 2016, pp. 13–16.

[53] B. Hammer, A. Schulz, and A. Gisbrecht, "Parametric nonlinear dimensionality reduction using kernel t-SNE reduction using kernel t-SNE," *Neurocomputing*, vol. 147, pp. 71–82, 2015.