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

Feature space visualization with spatial similarity maps for pathological speech data

Philipp Klumpp[1], *Juan Camilo Vásquez-Correa*[1,2], *Tino Haderlein*[1], *Elmar Nöth*[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]Universidad de Antioquia, Medellín, Colombia

{philipp.klumpp, juan.vasquez, tino.haderlein, elmar.noeth}@fau.de

Abstract

The feature vectors of a data set encode information about relations between speaker groups, clusters and outliers. Based on the assumption that these relations are conserved within the spatial properties of feature vectors, we introduce similarity maps to visualize consistencies and deviations in magnitude and orientation between two feature vectors. We also present an iterative approach to find subspaces of a high-dimensional feature space that encode information about predefined speaker clusters. The methods were evaluated with two different data sets, one from chronically hoarse speakers and a second one from Parkinson's Disease patients and a healthy control group. The results showed that similarity maps provide a decent visualization of speaker groups and the spatial properties of their respective feature vectors. With the iterative optimization, it was possible to find features that show pronounced spatial differences between predefined clusters.

Index Terms: feature space visualization, feature analysis, feature space reduction

1. Introduction

The first implementation step of any reliable classification algorithm is a good selection of features. One has to validate that the collected observations provide a meaningful information basis to solve a given problem. There are plenty of features that can be computed for a speech signal: ASR for instance makes use of short-time MFCCs as a representation of the power spectrum [1]. For emotion recognition, features like mean F_0 and F_0 variance have been successfully applied [2]. Other prosodic features based on intonation, loudness, speed and rhythm turned out to be useful for the evaluation of substitute voice after laryngectomy [3].

With the knowledge about the large variety of speech characteristics, we want to propose a visualization and feature analysis method that enables researchers to make intuitive decisions on whether a feature set is well-suited for a particular classification task or not. The method also provides information about how feature vectors behave with respect to different classes. Our approach is based on the assumption that all classes can be represented by different clusters in a high-dimensional feature space. Feature vectors are compared to each other with respect to two similarity metrics describing the differences in magnitude (L^2 norm) and orientation (cosine similarity). Both of these metrics have already been utilized for various classification tasks. A cosine similarity scoring introduced in [4] was successfully used in unsupervised speaker adaptation for speaker verification [5]. Apart from speech processing, classification based on cosine similarity yielded good results for face verification [6] and document clustering [7]. Another work on text document clustering incorporated both magnitude and orientation similarity metrics [8]. In the field of computer vision, Wang et al. proposed the Weighted Local Cosine Similarity as an enhancement of cosine similarity for visual tracking [9].

Similarity matrices have been used as a visualization tool for audio data since they were introduced in [10]. In the original application, they were utilized to compare an audio track with itself at certain time shifts. Our method makes use of similarity matrices to visualize differences between feature vectors with respect to the similarity metrics explained before. We want to show that for pathological speech data, similarity maps are particularly advantageous compared to other popular feature space visualizations (e.g. t-SNE [11], Diffusion maps [12], Isomaps [13]), because they allow the user to manually arrange feature vectors in the visualization by a parameter of their choice, for example a label or a patient property, such as age or gender. In the resulting plots, one can determine how similar several groups are to each other. With the proposed similarity maps, it is also possible to determine a small subset of features that best represents a predefined ideal similarity target through an optimization process.

After a brief explanation of the implementation of the proposed method, we evaluate the procedure with two independent data sets. We provide several exemplary visualization results we created with our approach and describe the benefits and limitations of our method in detail.

2. Materials and Methods

2.1. Computation of similarity maps

Given a set of observations in feature space of dimension n and their corresponding labels, we compared each vector with every other vector by computing two similarity measures. The results were stored in similarity matrices of size n^2. Afterwards, the columns and rows of each matrix were sorted with respect to the given label. For visualization, the sorted similarity matrices were plotted in a heat map.

For the implementation, we defined the magnitude difference between two observations by

$$\mathbf{M}_{\mathrm{mag}}[i,j] = \left| \|\overline{\mathbf{x}}_i\|_2 - \|\overline{\mathbf{x}}_j\|_2 \right| . \tag{1}$$

The cosine similarity metric was defined as

$$\mathbf{M}_{\cos}[i,j] = 1 - \frac{\overline{\mathbf{x}}_i \overline{\mathbf{x}}_\mathbf{j}}{\|\overline{\mathbf{x}}_i\|_2 \|\overline{\mathbf{x}}_j\|_2} . \tag{2}$$

Here, we subtracted the common cosine similarity formula from 1. The resulting output then ranged from 0 for identical angular orientation to 2 for opposing angular orientation in feature space, instead of [–1, 1]. As a result, for both the magnitude as well as the cosine metric, darker regions in the similarity map correspond to values close to 0 and indicate

http://dx.doi.org/10.21437/Interspeech.2019-2080

higher similarity between two observations.

2.2. Hoarse speech data set

We evaluated this visualization method with two data sets. The first one was collected from 73 German subjects with chronic hoarseness. Patients suffering from cancer were excluded. Each person read the text "Der Nordwind und die Sonne" ("The North Wind and the Sun", [14]), a phonetically rich standard text which is frequently used in clinical speech evaluation in German-speaking countries. It contains 108 words (71 distinct) with 172 syllables. Five voice professionals (one ear, nose and throat doctor, four speech therapists) evaluated the intelligibility of each recording perceptually. Rating was performed on a five-point Likert scale [15]. For computation of average scores for each patient, the grades were converted to integer values (1 = 'very high', 2 = 'rather high', 3 = 'medium', 4 = 'rather low', 5 = 'very low'). For each patient, an intelligibility mark, expressed as a floating point value, was calculated as the arithmetic mean of the single scores.
All recordings were then processed by a speech recognizer, which received minor adaptations to better match our requirements. The recognition vocabulary was changed to the 71 words of the standard text. Only a unigram language model was used so that the results mainly depended on the acoustic models. In order to find counterparts for intelligibility, a prosody module was used to compute features based upon frequency, duration, and speech energy (intensity) measures. The prosody module processed the output of the word recognition module and the speech signal itself. Originally, there were 95 local prosodic features computed for each word position. After several studies on voice and speech assessment, however, a relevant core set of 33 features had been defined for further processing [16], including statistics about pauses, energy, duration and F_0. The 33 local features per word were then averaged with respect to different conditions, for example over all words, over all nouns, or over all verbs and nouns. In addition to this, 15 global features were computed for intervals of 15 words length each. They covered the means and standard deviations of jitter and shimmer, the number, length, and maximum length of voiced and unvoiced sections, the ratio of voiced and unvoiced sections, the ratio of the length of the voiced sections to the length of the signal, and the same for unvoiced sections [17]. The last feature was the standard deviation of F_0. The final feature vector contained a total of 283 features, and every dimension of it was scaled to the interval [–1, 1]. The human listeners rated with respect to the entire text. In order to receive one single value for each feature per speaker, the average of each prosodic feature over all selected words served as the final observation value.

2.3. Parkinson's Disease data set

Additionally, we conducted a second evaluation with a data set collected from 85 participants, 37 of them diagnosed with Parkinson's Disease (PD). This is a subset of the database used in [18]. Within the PD group, there were 22 female and 15 male participants. The healthy control group was composed of 25 female and 23 male speakers, respectively. Every PD patient was examined based on the Unified Parkinson's Disease Rating Scale (UPDRS) [19], where higher scores indicate a higher progression of the disease. The distribution of UPDRS scores for the PD group is shown in Figure 1, for the healthy controls (HC), the score was assumed to be 0. The participants were

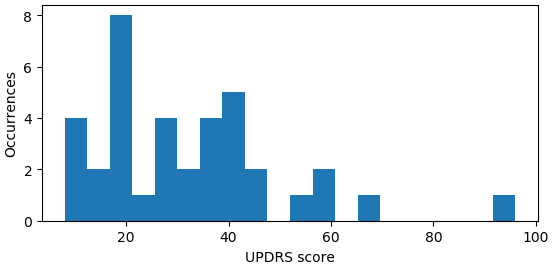

Figure 1: *Distribution of UPDRS scores of all Parkinson's Disease subjects.*

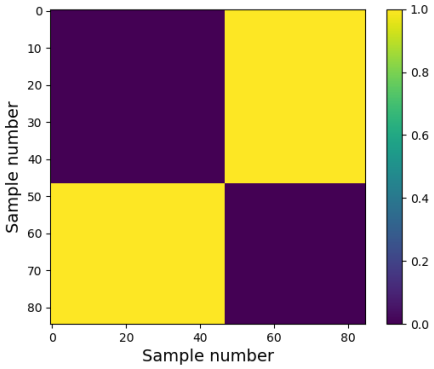

Figure 2: *Target similarity map for the optimization process. Top left (female) and bottom right (male) represented the samples of the same gender.*

asked to give a short monologue (1 minute) about their everyday routine. 38 prosodic features were extracted with an extended version of Neurospeech[1] [20], covering statistics about F_0, energy, voiced or unvoiced segments and pauses. Every feature was scaled to a range of [–1, 1]. For the computation of the cosine similarity of the PD data set, we also removed the mean from all features to demonstrate how this would affect the results.

2.4. Evaluation and optimization of similarity maps

For our experiments, we visualized the similarity maps for the full feature vectors as well as for a manually reduced feature space and analyzed how specific subsets influenced the similarity results for hoarse speakers. For the PD data set, we improved the method by applying an iterative optimization approach with predefined target maps. The algorithm iterated over all possible subsets of feature vectors for a defined subset size $n < N$. Note that this approach is only feasible for small subset sizes, because the number of possible combinations Ω rapidly increases for greater values for n:

$$\Omega = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{3}$$

For each combination, the cosine and magnitude similarity maps were computed. Afterwards, the absolute error between

[1] https://github.com/jcvasquezc/DisVoice

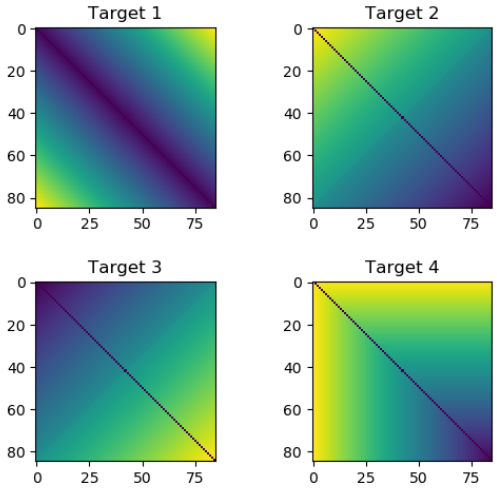

Figure 3: *Different target similarity maps which can be used for the optimization process of feature vectors sorted according to a continuous label. Target 1: Closer feature vectors indicate closer labels. Target 2: Increased similarity with increased labels. Target 3: Decreased similarity with increased labels. Target 4: Similar to Target 2, but pronounced dissimilarity between two distant labeled feature vectors.*

each map and the target map was computed, and the ten combinations with the smallest error were returned for each target. The design of a particular target depends on the number of samples for each class and on the desired similarity of the feature subset. Figure 2 shows a target map that was used to find a reliable subset of features that differ between genders. It can be seen that we expected high similarities (small values in the map) within the groups of female (top left) and male (bottom right) speakers, and low similarities when comparing samples from opposing groups (top right and bottom left). The optimization algorithm is independent of the data, but it depends on the target property. Unlike for discrete classes, in the continuous case it is necessary to provide continuous target maps, which ideally incorporate prior knowledge about the expected similarity distribution. Four exemplary targets are shown in Figure 3.

3. Results

Figure 4 shows the cosine similarity map for the full 283-dimensional feature vectors of the hoarse speakers. For all visualizations of the hoarse speech data set, files were ordered with respect to the intelligibility rating given by voice professionals. Lower sample numbers corresponded to a higher (better) intelligibility rating.

In the cosine similarity map, we could see a clear trend. The top left region, in which highly intelligible speakers were compared to each other, was darker than the rest of the map. In the regions in the top right and bottom left, where highly intelligible speakers were compared with poorly intelligible ones, the difference indicated by the similarity map was very high. Along the main diagonal, where we compared speakers with similar intelligibility rating, the similarity matrix indicated a slight increase of deviation in the direction from top left to bottom right. The maximum cosine deviation was almost 0.3.

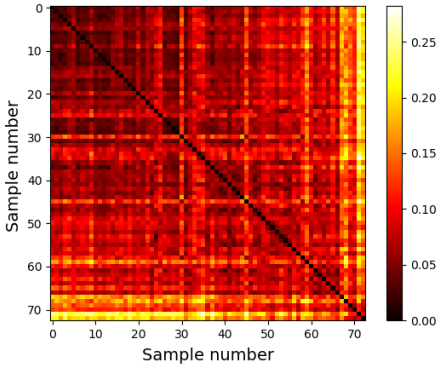

Figure 4: *Cosine similarity map for the full feature space of the hoarse speech data set. Smaller sample numbers correspond to higher intelligibility rating.*

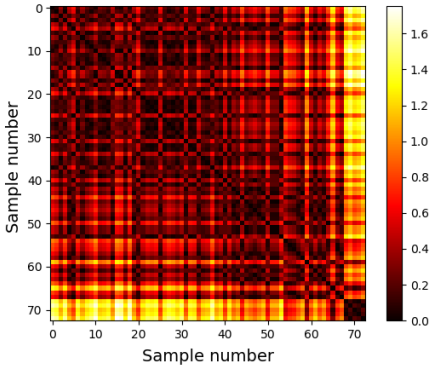

Figure 5: *Magnitude similarity map for the hoarse speech data set with reduced feature space. Smaller sample numbers correspond to higher intelligibility rating.*

Afterwards, we performed a manual selection based on our experience with hoarse speech, extracting 20 features from the 283-dimensional feature vector (e.g. word accuracy, mean F_0 and jitter) that we expected to encode rich information about the degree of hoarseness for a speaker. This resulted in a reduction of the feature space by approximately 93%. For this smaller feature vector, we computed a magnitude similarity map which is depicted in Figure 5. Looking at the bottom left and top right regions, where we compared speakers with a high difference in their respective intelligibility rating, this difference was also present in the similarity map. The five participants with the lowest intelligibility score showed a high dissimilarity with all other speakers, but were quite similar among their small group. For the PD data, the full feature space was reduced to subsets by iterative optimization with different target similarity maps. To verify the reliability of the method, we conducted a first optimization searching for subsets of two features that were the most gender-dependent, resulting in the set of average fundamental frequency in voiced segments and voiced duration regularity. The samples in the respective cosine similarity map (Figure 6) were sorted by gender (female to male), and within the same gender by increasing UPDRS score. Note that for this similarity map, the values ranged from 0 to 2 after the mean

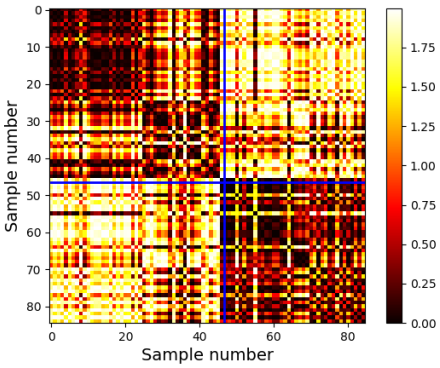

Figure 6: *Cosine similarity map for the PD speech data set with only two features, optimized for gender identification. The blue bound separates both genders: Top left female, bottom right male samples.*

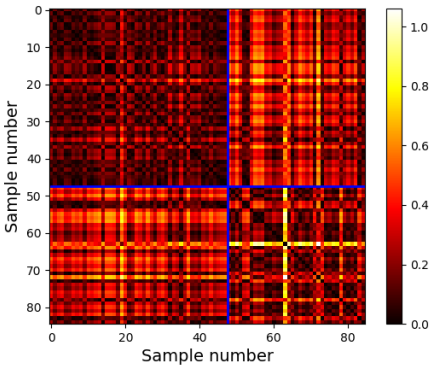

Figure 7: *Magnitude similarity map for the PD speech data set with four features, optimized for PD identification. The blue bound separates both groups: Top left HC, bottom right PD samples.*

of each feature has been normalized to 0 for the cosine metric only. It can be seen that the similarity between samples of the same gender is higher than the one for opposing genders. Within the cluster of the female speakers, we found a second cluster formed by female patients with further progression of the disease. There were also a couple of outliers for both genders.

In the last step, we ran a second optimization, ordering the samples only by their UPDRS score from smallest to highest, to find a set of four features which would best separate HC from PD participants. The resulting features were average fundamental frequency in voiced segments, standard deviation of energy, ratio of voiced to silence duration and voiced duration regularity. Figure 7 shows the magnitude similarity for the 4-features set. The HC group yielded a high similarity within itself, and clearly separated from the PD samples. The PD patients showed a lower similarity within their group, and there is one pronounced outlier for sample number 63.

4. Discussion

For the hoarse speech data, we could show that the decrease in intelligibility affected the spatial properties of feature vectors. The more affected the speech of a participant was, the higher was the dissimilarity of the sample with samples from less affected speakers. This trend became a lot more pronounced after reducing the feature space in the cosine similarity map. The five least intelligible speakers formed a cluster in the bottom right corner, indicating that the speech of these patients became more similar with respect to certain features and it clearly separated them from their healthier counterparts.

With the iterative optimization, we managed to find two features that clearly separated female and male speakers. The best ten results for the target map introduced before all contained the average fundamental frequency in voiced segments as a good feature. This outcome was expected and it showed that the optimization yielded appropriate feature subsets. Overall, the cosine similarity map in Figure 6 showed a high difference in the angular orientation of feature vectors. This was caused by the normalization to zero mean. We want to point out that a zero mean normalization can be helpful to pronounce differences in cosine similarity maps. However, it is important to keep in mind that zero mean normalization transforms magnitude dissimilarities to opposing angular orientations in the cosine map. An investigation of the outliers within each cluster showed that they were caused by female speakers with a relatively low and male speakers with a relatively high F_0.

The optimization results for the separation of HC and PD samples yielded a significant magnitude map which showed how similar the computed four-features-set performed for the HC group. At the same time, we could observe a clearly visible difference when we compared two samples from opposing groups. The pronounced outlier was caused by a very high ratio of voiced to silence duration. An investigation of the recording showed that the subjects speech was fluent and with minimal pauses, unlike many of the other samples.

5. Conclusion

Cosine and magnitude similarity maps turned out to be a decent tool for the visualization of feature spaces. Magnitude and angular orientation of feature vectors encode properties of different clusters within a data set. These spatial differences can be used to identify features that are relevant for a specific classification task. In the homogeneous clusters of a similarity map, outliers become clearly visible. Furthermore, it is possible to track down the features that cause the outlier and give an explanation for the observed deviation. With the iterative optimization method, it is also possible to find features that contribute to predefined spatial transformations of feature vectors. Most importantly, similarity maps help to better understand how certain features or a small subset of them change for different speaker groups, and they enable us to reasonably explain outliers.

6. Acknowledgements

The work reported here was financed by CODI from University of Antioquia by grants Number 2015–7683. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287.

7. References

[1] M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, pp. 763–786, 2007.

[2] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 157–183, 2003.

[3] T. Haderlein, E. Nöth, H. Toy, A. Batliner, M. Schuster, U. Eysholdt, J. Hornegger, and F. Rosanowski, "Automatic evaluation of prosodic features of tracheoesophageal substitute voice," *European Archives of Oto-Rhino-Laryngology*, vol. 264, no. 11, pp. 1315–1321, 2007.

[4] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification," Ph.D. dissertation, École de technologie supérieure, Montréal, Canada, 2009.

[5] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification." in *Odyssey*, 2010, p. 16.

[6] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 709–720.

[7] L. Muflikhah and B. Baharudin, "Document clustering using concept space and cosine similarity measurement," in *International Conference on Computer Technology and Development, 2009. ICCTD'09.*, vol. 1. IEEE, 2009, pp. 58–62.

[8] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.

[9] D. Wang, H. Lu, and C. Bo, "Visual tracking via weighted local cosine similarity," *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1838–1850, 2015.

[10] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, ser. MULTIMEDIA '99. New York, NY, USA: ACM, 1999, pp. 77–80. [Online]. Available: http://doi.acm.org/10.1145/319463.319472

[11] 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.

[12] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[13] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[14] International Phonetic Association (IPA), "Handbook of the International Phonetic Association," Cambridge University Press, Cambridge, 1999.

[15] R. Likert, "A technique for the measurement of attitudes." *Archives of Psychology*, vol. 140, 1932.

[16] T. Haderlein, C. Schwemmle, M. Döllinger, V. Matoušek, M. Ptok, and E. Nöth, "Automatic Evaluation of Voice Quality Using Text-based Laryngograph Measurements and Prosodic Analysis," *Computational and Mathematical Methods in Medicine*, vol. 2015, 2015, 11 pages. Published June 2, 2015.

[17] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The Prosody Module," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin: Springer, 2000, pp. 106–121.

[18] J. C. Vasquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. M. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: a deep learning approach," *IEEE Journal of Biomedical and Health Informatics*, 2018.

[19] C. G. Goetz, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, G. T. Stebbins, M. B. Stern, B. C. Tilley, R. Dodel, B. Dubois *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, no. 1, pp. 41–47, 2007.

[20] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei *et al.*, "Neurospeech: An open-source software for Parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, 2018.