smc18
sonic crossings

15th Sound & Music Computing Conference

Sonic Crossings

4–7 July
Limassol, Cyprus

PAPER PROCEEDINGS

Proceedings of the 15th Sound and Music Computing Conference (SMC 2018), Limassol Cyprus, 2018

Editors

Dr. Anastasia Georgaki and Dr. Areti Andreopoulou

ISBN

978-9963-697-30-4

Relative DTW embedding for binary classification of audio data

Marcelo Queiroz
University of São Paulo
mqz@ime.usp.br

Guilherme Jun Yoshimura
University of São Paulo
jun@ime.usp.br

ABSTRACT

This paper presents a novel classification technique for datasets of similar audio fragments with different durations, that allows testing pertinence of fragments without the need of embedding data in a common representation space. This geometrically-motivated technique considers direct DTW measurements between alternative different-sized representations, such as MFCCgrams or chromagrams, and defines the classification problem over relative embeddings based on point-to-set distances. The proposed technique is applied to the classification of voice recordings containing both normal and disturbed speech utterances, showing that it significantly improves performance metrics with respect to usual alternatives for this type of classification, such as bag-of-words histograms and Hidden-Markov Models. An experiment was conducted using the Universal Access Speech database (UA-Speech) from the University of Illinois, which contains over 700 different words recorded by 19 dysarthric speakers and 13 speakers without any speech disorder. The method proposed here achieved a global F-measure (with 10-fold cross-validation) above 95%, against 81% for a bag-of-words classification and 83% for Hidden Markov Models.

1. INTRODUCTION

There are many important practical problems involving classes of similar audio items with different durations, e.g. the development of voice command devices, where user commands have the form of vocal utterances with varying speed and timbre [1, 2], query-by-humming mechanisms, where database queries have the form of sung melodies to be matched against symbolic data [3], synchronization of different music performances [4], among many others. In all these cases classification problems may have to be solved somewhere in the processing chain: which among a set of available voice command is this utterance most similar? Which known melodies have similar interval profiles? Is this audio recording a potential reinterpretation of a given song?

Usually these problems are recast in terms of comparing matricial data in frame-based feature spaces which are known to provide meaningful alternative representations for the original audio data, as MFCCgrams for speech utterances and f0-grams or chromagrams for melodies and harmonies. Such frame-based feature spaces are heterogeneous spaces in the sense that they contain matrices of different lengths [1] and for this reason items cannot be easily combined using the regular linear-algebraic DSP artillery (consider trying to take the centroid of several MFCCgrams of different lengths). Figure 1 illustrates both the heterogeneous nature of the frame-based feature space as well as the motivation for the relative embedding here proposed, which is to allow viewing a certain class of heterogeneous items through a simple geometric model.

The embedding of the original data in such heterogeneous frame-based feature spaces allows direct comparison of pairs of items, for instance using Dynamic Time Warping [5] as a means of establishing similarity between data of differing lengths, and this synchronization mechanism may even entail strategies for alleviating the lack of said linear operators (e.g. temporally synchronizing data before taking averages), but it still does not allow simple characterizations of sets of several similar items, such as Gaussian models based on centroids and covariance matrices.

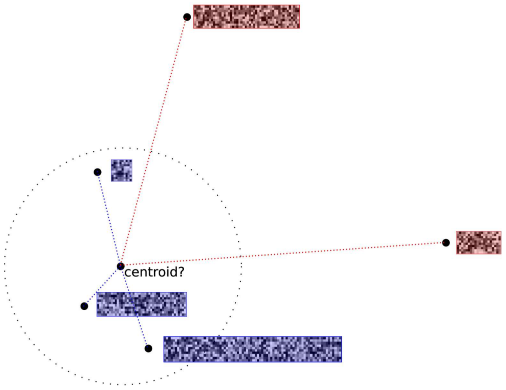


Figure 1. Binary classification setup with heterogeneous items. Items may be MFCCgrams or Chromagrams, or any other type of frame-based feature matrix.

There are a few off-the-shelf strategies that allow embedding of such items in homogeneous metric spaces, over which machine learning algorithms can be trained: in very simple lines, a first approach consists in applying Vector Quantization (VQ) to the frame-dependent features vectors followed by Bag-of-Words (BoW) modeling [6], a sec-

[1] We will use the term *length* to refer to the temporal dimension of the frame-based feature matrix, i.e. length is the number of frames used to segment the audio.

ond approach uses Hidden Markov Models (HMM) [7]. VQ+BoW starts with a clusterization (e.g. using Kmeans) of all features known to a training stage, using K clusters, and then each item is encoded in a K-dimensional histogram, where each counter $k = 1, 2, \ldots, K$ represents how many of this item's feature vectors belong to the k-th cluster; in this way all variable-length items are represented through homogeneous (and possibly length-normalized) K-dimensional features, allowing all sorts of operations to be performed that are useful for classification, including the use of Support Vector Machines (SVM) [8]. With HMM a very different sort of implicit representation is built, by viewing in-class heterogeneous items as observations produced with high probability by a Markov model, where K interconnected inner states reflect the temporal/stochastic evolution of data, and feature vector emission probabilities provide the statistical link between inner states and observed feature vectors; classification then proceeds by Viterbi reconstruction of optimal paths through the Markov chain, with associated probabilities that allow discrimination between in-class and out-of-class items.

Both VQ+BoW and HMM have some drawbacks. VQ+BoW throws away the temporal sequence of features: essentially any reordering of the same frame-based feature matrix would produce the same histograms. This may represent a huge problem or no problem at all, according to the application context: models for global timbre [9], chord recognition [10] or speaker identity [11] are usually based on frame-based features without temporal nexus, but word recognition [7], melody recognition [3], and cover-song identification [12] are all time-dependent tasks, for which reordering of frames makes absolutely no sense. HMM preserves the temporal nexus of the frame-based feature matrix, but length differences have an impact on the Viterbi reconstructions in the sense that spending too many frames on a single node produces a decreased probability with respect to a similar signal that moves on through the Markov chain at a quicker pace, i.e. an HMM model penalizes lengthier data with respect to shorter data.

It should by now be also immediate that the use of well-known classifiers such as SVMs *within the original representation space* is not possible without a prior embedding into a homogeneous feature space, e.g. via VQ+BoW or HMM. Taking Figure 1 again as an example, an SVM approach would require not only this prior embedding, but also the definition of a kernel function in order to bend the linear classifier around a more or less rounded class, an operation which would also depend on the notion of a centroid, absent in our original heterogeneous representation space.

The goal of this paper is to propose a classification strategy for time-dependent audio data that extends DTW distances [2] between pairs of heterogeneous frame-based feature matrices to point-to-set distances that allow a relative embedding of multidimensional data into a relative distance space used for actual classification. This embedding is relative because items have no fixed position, in fact items are positioned only relatively to the whole set of in-class items. The motivation is to provide a surrogate representation for the centroid+radius idea of Figure 1 that does away with the need of a centroid and yet allows the classification based on a simple distance-based geometric criterion. The proposed strategy is then applied to a classification problem related to disturbed versus regular speech, and compared to both VQ+BoW and HMM approaches.

The structure of the paper is as follows. Section 2 formalizes the proposed relative DTW embeddings and its corresponding classification strategy. Section 3 presents the experimental methodology for the application of the proposed strategy to the speech classification problem, whose results are then presented and discussed in Section 4. Concluding remarks and further work are outlined in Section 5.

2. DTW RELATIVE EMBEDDINGS

As suggested in the Introduction, the motivation for the strategy to be presented is allowing us to look at a class of heterogeneous time-dependent audio data using a Gaussian-like geometric model, through a relative embedding that considers point-to-set distances from items to the modelled class (see Figure 1).

Consider a set of items or frame-based feature matrices $\mathcal{N} = \{M^0, M^1, \ldots, M^N\}$ that comprise the class of interest for the classification problem. We will extend the regular DTW distance between items to allow computation of *point-to-class* distances for any item within this heterogeneous feature space. Specifically, let x be an arbitrary (in-class or out-of-class) item within the feature space, and let

$$\mathrm{MinDTW}(x) = \min_{y \in \mathcal{N} \setminus \{x\}} \mathrm{DTW}(x, z), \tag{1}$$

be the smallest DTW distance from x to any (other) in-class item y, i.e., MinDTW(x) expresses how close is x to the closest representative of class $\mathcal{N}$ that is not x itself. The mapping $x \mapsto \mathrm{MinDTW}(x)$ is named *relative DTW embedding* of x since it does not position items in any absolute manner, but simply places them in a one-dimensional space relatively to class $\mathcal{N}$.

Take for instance the example in Figure 2, where a class $\mathcal{N}$ consists of 3 blue items, and there are 2 red out-of-class items. For each item x, MinDTW(x) is represented by an outbound arrow departing from x and reaching the representative $y \in \mathcal{N} \setminus \{x\}$ closest to x. In this case the two lower in-class items are very close to each other and their MinDTW(x) values are the same; a borderline in-class item stands relatively farther away, and out-of-class items are reachable through a longer path. It should be noted that distances between out-of-class items are not used in any way in the embedding, and also that when new items are included in the class, all values MinDTW(x) either lower or stay the same (by monotonicity of the $\min$ operator with respect to set inclusion).

The relative DTW embedding is defined for all items in the heterogeneous feature space, and its usefulness is dependent upon a property of class $\mathcal{N}$, namely that in-class

[2] It is well-known that DTW, as the cosine distance, is not formally a distance due to its violation of the triangle inequality. We will however stick to the MIR tradition of referring to these dissimilarity measures or pre-metrics as distances.

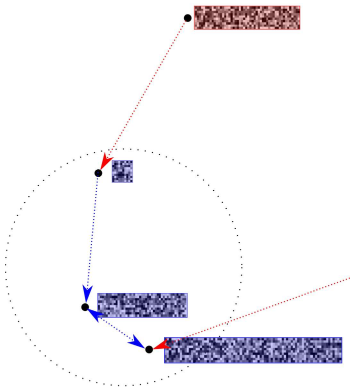


Figure 2. Binary classification setup using relative DTW embedding. Item x is associated to its distance to the closest representative $y \neq x$ within class $\mathcal{N}$.

items are positioned closely together with respect to out-of-class items. In other words, it is assumed, for the purpose of applicability of this model, that intra-class distances $\mathrm{DTW}(x, y)$ for $x, y \in \mathcal{N}$ are generally smaller than distances $\mathrm{DTW}(x, y)$ between in-class and out-of-class items ($x \in \mathcal{N}$, $y \notin \mathcal{N}$), or equivalently that the statistical distributions of intra-class distances and in-class/out-of-class distances are significantly different. This assumption should hold for many of the application examples mentioned in Section 1, e.g. cover-song identification, melody matching in query-by-humming, and disturbed speech classification (the latter is addressed in Sections 3 and 4).

Based on the above assumption, a simple classification strategy is defined by means of characterizing the borders of the in-class and out-of-class items. Specifically, let

$$\varrho^+ = \max_{z \in \mathcal{N}} \mathrm{MinDTW}(z) \tag{2}$$

and

$$\varrho^- = \min_{z \notin \mathcal{N}} \mathrm{MinDTW}(z) \tag{3}$$

be the largest intra-class distance and the smallest out-of-class distance to the class. If it should happen that $\varrho^+ < \varrho^-$ then perfect separation between in-class and out-of-class items is achieved, and an intermediary threshold such as

$$\tau = \frac{\varrho^+ + \varrho^-}{2} \tag{4}$$

may be used for classification of new unknown items, according to

$$\begin{cases} z \in \mathcal{N} & \text{if } \mathrm{MinDTW}(z) < \tau \\ z \notin \mathcal{N} & \text{otherwise.} \end{cases} \tag{5}$$

In general it could happen that the relative DTW embeddings of in-class and out-of-class are not perfectly separable (i.e. $\varrho^+ \geq \varrho^-$), and then a more fitting threshold may be defined by taking the optimal value $\tau \in [\varrho^-, \varrho^+]$ according to a performance measure of interest, e.g.

$$\tau = \operatorname*{argmax}_{\alpha \in [\varrho^-, \varrho^+]} \text{F-measure}(\alpha), \tag{6}$$

where the F-measure is computed by applying the above classification strategy to all known labeled items available to be used during the training stage.

There are other possibilities for defining similar relative embeddings of items from a heterogeneous frame-based feature space to a unidimensional point-to-class relative distance space, using DTW as a means to preserve time-coherency of the relative measurements. One such alternative is the use of the Hausdorff distance H, defined for general sets A, B and any given distance d as

$$H(A, B) = \max\left(\sup_{x \in A} \inf_{y \in B} d(x, y), \inf_{x \in A} \sup_{y \in B} d(x, y) \right), \tag{7}$$

i.e. the distance between the sets is the largest distance you are forced to travel from some point of one set to the closest point of the other set. When one of the sets is unitary, the above expression simplifies to $H(x, B) = \sup_{y \in B} d(x, y)$, from which we define a relative DTW embedding as

$$\mathrm{HausdorffDTW}(x) = \max_{y \in \mathcal{N} \setminus \{x\}} \mathrm{DTW}(x, y). \tag{8}$$

The main motivation for considering Hausdorff distances in this classification context is the fact that out-of-class items are now being compared to the farthest in-class item possible, which might make the classification task easier. It is also true however that intra-class distances will in general increase, but by how much they will increase depends on the distribution of the values $\mathrm{DTW}(x, y)$ for $x, y \in \mathcal{N}$, e.g. if all values are very close (not necessarily close to zero) as typically occurs for time-warped versions of the same signal, then the relative HausdorffDTW embeddings of the in-class items could remain more or less in the same region. It should be noted that, since DTW does not satisfy the triangle inequality, it is not necessarily true that when $\mathrm{DTW}(x, y)$ are small for $x, y \in \mathcal{N}$ then the values $\mathrm{HausdorffDTW}(x, w)$ and $\mathrm{MinDTW}(x, w)$ would be close for $x \in \mathcal{N}$ and $w \notin \mathcal{N}$.

3. EXPERIMENTAL METHODOLOGY

We will illustrate the proposed classification strategy based on relative DTW embedding by employing it in a disturbed speech classification problem. In this problem, the goal is to classify a speech recording, usually of a specific word, as normal or disturbed based on its similarity to a set of reference speech recordings. This classification problem is asymmetric in the sense that the class of normal utterances of a given word are relatively homogeneous and using DTW to compare any pair of normal utterances typically produces small values, but disturbed utterances have no intrinsic similarity to each other, due to the large number of different speech disorders that cause them, and also due to differences in the syllable where a deviation is observed.

In this experiment we use the Universal Access Speech database (UA-Speech) recorded at University of Illinois by Mark Hasegawa-Johnson's group [13], which is one of the largest databases available for disordered speech. It contains speech recordings by 19 speakers with cerebral palsy, and reference recordings of normal speech by 13 speakers, each one contributing with a total of 765 recordings of isolated words, separated according to word categories, which are: 10 digits (e.g. 'Zero' to 'Nine'), 26 NATO phonetic alphabet words (e.g. 'Echo', 'Sierra', 'Bravo'), 19 computer commands (e.g. 'Delete', 'Enter', 'Command'), 100 common words (e.g. 'The', 'It', 'In') and 300 uncommon words (e.g. 'Naturalization', 'Hypothesis'). Each speaker has recorded each word 3 times, with exception of uncommon words which have a single recording, and each recording has been captured by 8 microphones. We will refer to as *UA-Speech Original* the dataset using only the first microphone per recording, and as *UA-Speech Extended* the dataset with all 8 microphones. In addition, the UA-Speech dataset contains speaker details such as age, sex, range of intelligibility (very low, low, mid, high) and type of dysarthria. Recordings are encoded as mono wav type audio files with 48kHz sampling rate.

Figure 3 shows the required steps to classify a new speech recording, i.e. feature extraction, similarity measurements, and proper classification, which will be detailed below.

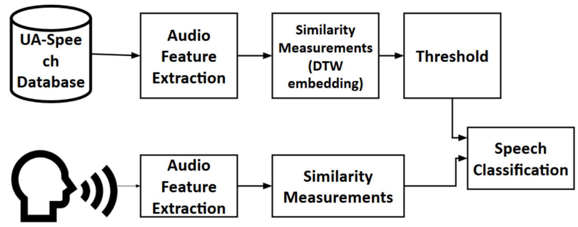


Figure 3. Stages of the speech classification experiment. The upper row represents the training stage from labeled data in UA-Speech database, and the lower row stands for the classification of new items. In a cross-validation context, a selection of the database is used for training and the remaining items are used for testing.

In the audio feature extraction phase we use Mel Frequency Cepstrum Coefficients (MFCC) to represent each audio frame; MFCCgrams are the preferred frame-based matrix representation for phoneme-related speech processing [14], and this is especially true in the speech classification problem here considered, because dysarthric utterances are essentially modifications of the phonetic contents with respect to normal utterances, which are otherwise very similar in terms of other audio characteristics [15]. Each MFCC vector with 12 coefficients is obtained from audio frames of 2048 samples (with 75% overlap) using the librosa library [11] with parameters y = audio signal and sr = sample rate of the audio. Each speech recording i is then represented through its MFCCgram M^i

Based on all data available for training, we build a similarity matrix S from the DTW distances between all pairs of MFCCgrams:

$$S_{ij} = \mathrm{DTW}(M^i, M^j) \tag{9}$$

Figure 4 shows a similarity matrix for recordings of the word "SEVEN", where it is possible to see that the set of normal recordings $\mathcal{N}$ form a tightly connected cluster (small intra-class distances) whereas disturbed recordings are widely spread with respect to both normal and other disturbed recordings.

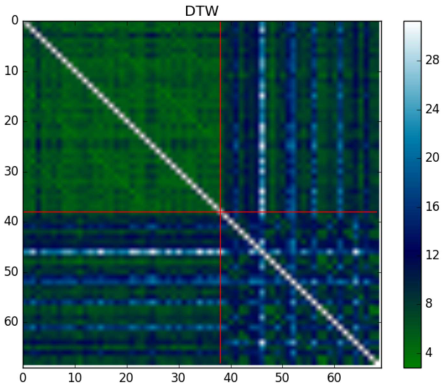


Figure 4. Similarity matrix for all recordings of the word "SEVEN". $\mathcal{N} = \{0, \ldots, 39\}$ correspond to the set of normal recordings in this example, and the remaining recordings present some form of disturb.

According to the relative DTW embedding strategy defined in Section 2, the set of normal recordings is used to define a classification threshold. For all recordings, we define the point-to-set distances:

$$\mathrm{MinDTW}(i) = \min_{j \in \mathcal{N} \setminus \{i\}} S_{ij} \tag{10}$$

$$\mathrm{HausdorffDTW}(i) = \max_{j \in \mathcal{N} \setminus \{i\}} S_{ij} \tag{11}$$

Classification is then done by sieving one of the above metrics through a simple threshold. This threshold is defined with respect to the relative diameter of the normal set, defined by the largest intra-class distance

$$\varrho^{+}(\mathrm{AnyDTW}) = \max_{i \in \mathcal{N}} \mathrm{AnyDTW}(i) \tag{12}$$

and the Smallest Out-of-class Distance

$$\varrho^{-}(\mathrm{AnyDTW}) = \min_{i \notin \mathcal{N}} \mathrm{AnyDTW}(i), \tag{13}$$

where AnyDTW stands for either MinDTW or HausdorffDTW. The threshold is defined as

$$\tau = \begin{cases} \dfrac{\varrho^{+} + \varrho^{-}}{2}, & \text{if } \varrho^{+} < \varrho^{-}, \\ \underset{\alpha \in [\varrho^{-}, \varrho^{+}]}{\mathrm{argmax}} \ \text{F-measure}(\alpha), & \text{otherwise.} \end{cases} \tag{14}$$

where F-measure(α) is computed with respect to all data available to the training stage.

We will compare our method to two other well-known methods: VQ+BoW and HMM. In VQ+BoW we first clusterize all MFCCs available for training in order to build normalized histograms to represent each recording; K-means is used for VQ with an optimal K chosen according to the silhouette coefficient. After that, the set of normal recordings is represented through a centroid of the normal histograms and a radius associated to a threshold in between the maximum distance from any normal histogram to the centroid, and the minimum distance from any disturbed histogram to the centroid. In HMM, the number K of inner states has been allowed to vary between 20 and 70; for each K, all training MFCCgrams are used to calibrate the transition matrix and emission probabilities, and then Viterbi probabilities are obtained for all normal and disturbed training data, and a threshold is selected in between the minimum normal speech probability and the maximum disturbed speech probability.

All methods are trained and tested following the same procedure, according to the same general guidelines:

- Each word in the dataset defines a different (DTW, BoW or HMM) model. Different words define different classification problems.
- All thresholds have been selected in order to optimize performance independently for each method.
- All methods are submitted to 10-fold cross-validation. For each fold k, 90% of data is used for training and 10% for testing, and the corresponding amounts of true positives TP[k], true negatives TN[k], false positives FP[k] and false negatives FN[k] are stored.
- Performance is evaluated using a global F-measure, defined as

$$F_{\text{global}} = \frac{2 \cdot \sum_k TP[k]}{2 \cdot \sum_k TP[k] + \sum_k FP[k] + \sum_k FN[k]}. \quad (15)$$

  This is done to minimize several types of biases associated to combining F-measure, Precision or Recall values from different folds [16].

4. RESULTS AND DISCUSSION

Figures 5 and 6 illustrate the result of one cross-validation fold for classification of the words "Command" (one of the computer command words) and "Hypothesis" (one of the uncommon words) using DTW relative embedding. The horizontal axis represent the relative embedding of each item with respect to the normal recordings used in training, and the vertical axis represent the values of the probability density function (pdf), blue for the datas labeled as normal and red for the datas labeled as disturb. Training data is identified by circles and squares, and test data by triangles.

In the particular fold displayed for the word "Command" (Figure 5) it can be seen that even though separation was not perfect for training data ($\varrho^+ > \varrho^-$) no test data appeared on the wrong side of the threshold, and so in this particular example the classifier got a perfect score (F-measure of 100%). In Figure 6 (example for the word "Hypothesis") it can be seen that a normal testing item was embedded to the right of some disturbed recordings, and the F-measure for this particular fold is 66%.

In Figures 7 and 8 it is possible to see that the distribution of normal and disturbed data using VQ+BoW is very overlapped, making it difficult to separate the classes. For the word "Command" using VQ+BoW the F-measure is 90% and for the word "Hypothesis" the F-measure is 80%.

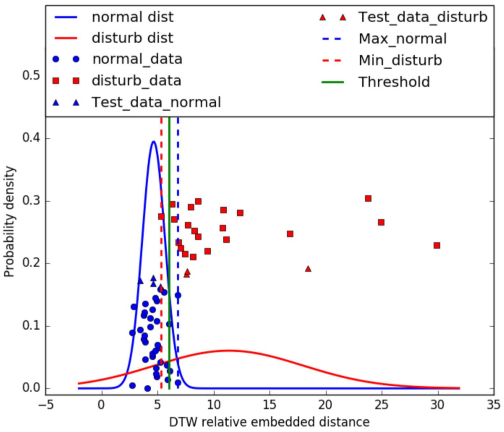


Figure 5. Classification of the word "Command" using DTW relative embedding. Circles and squares represent training data and triangles are testing data; horizontal axis is relative distance to the normal class; vertical lines represent (from left to right) ϱ^-, τ and ϱ^+.

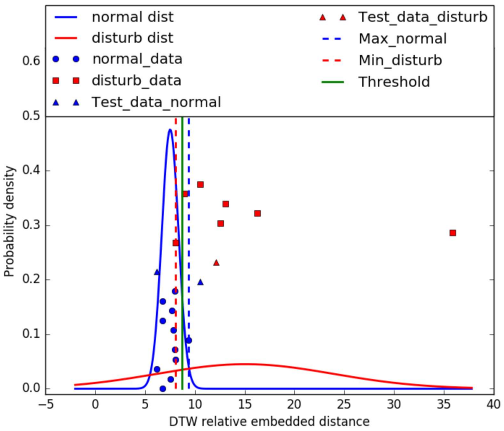


Figure 6. Classification of the word "Hypothesis" using DTW relative embedding.

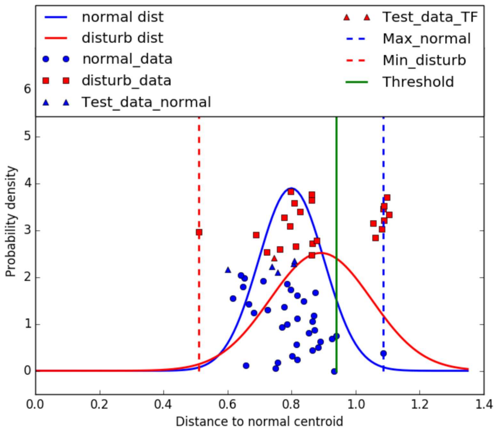


Figure 7. Classification of the word "Command" using VQ+BoW; horizontal axis is distance to the centroid of the normal class.

Figures 9 and 10 show the same example for HMM. Here again classes overlap, and for the word "Command" the

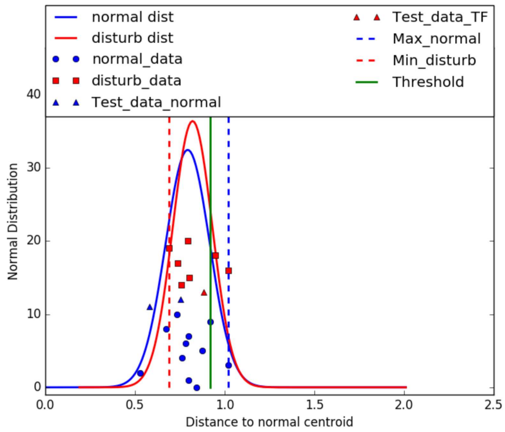


Figure 8. Classification of the word "Hypothesis" using VQ+BoW.

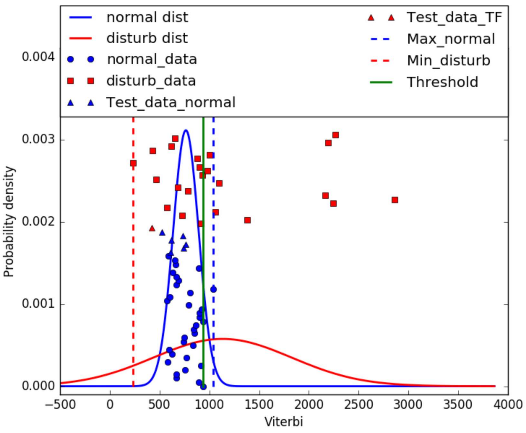


Figure 9. Classification of the word "Command" using HMM; horizontal axis is the absolute value of the Viterbi log-probability.

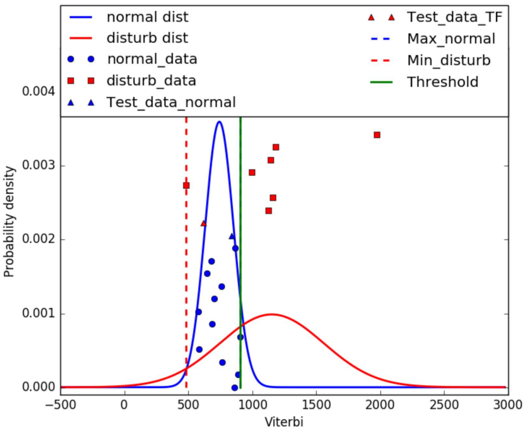


Figure 10. Classification of the word "Hypothesis" using HMM.

F-measure is 91% whereas for the word "Hypothesis" the F-measure is 66%.

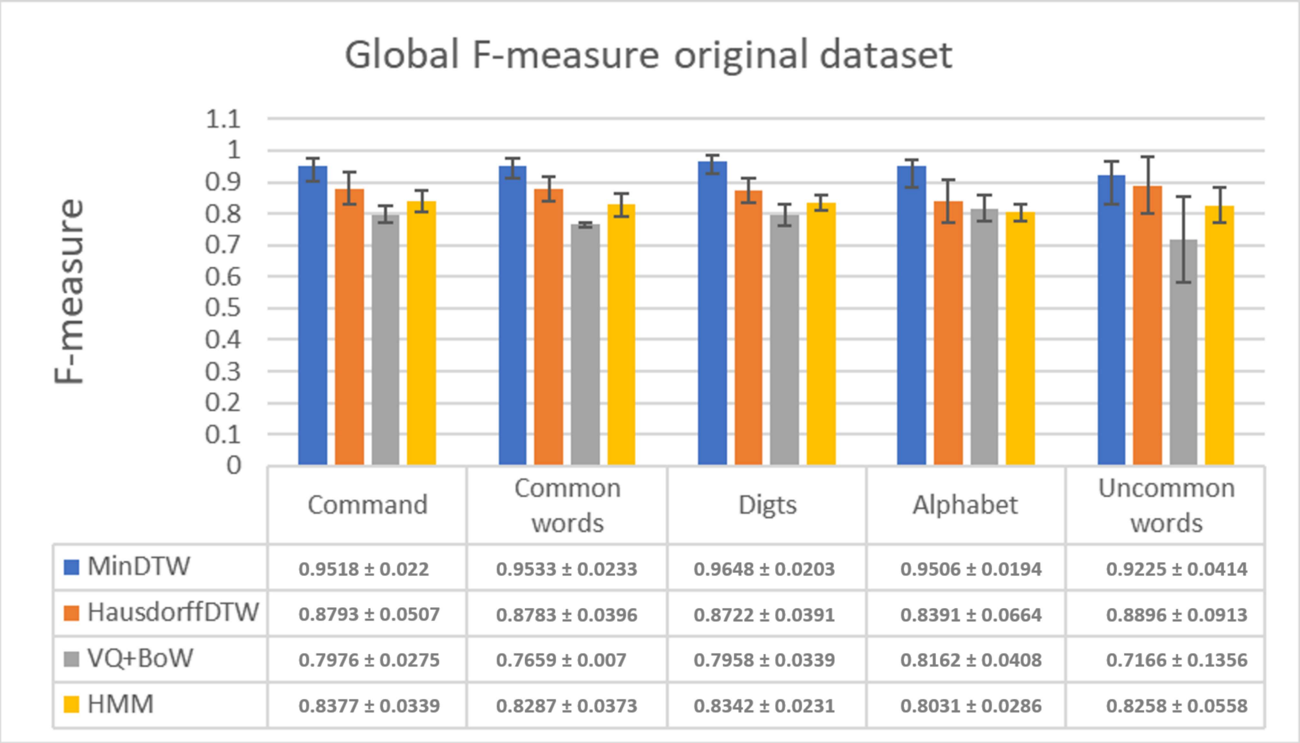
Figure 11 show comparative performance measurements (mean and variance of global F-measures) of the relative DTW embeddings and also VQ+BoW and HMM methods for all categories of words in the UA-Speech database, using the original dataset (single microphone). It is immediately noticeable that the proposed strategy is a huge improvement over the other two approaches for this dataset. Overall, HMM has better performance than VQ+BoW, which is expected, since this type of classification is sensitive to the temporal sequence of frames, an important aspect of time-dependent models such as HMM and DTW but is utterly ignored by the bag-of-words approach.

It can be seen in the MinDTW column of Figure11 that the uncommon words have the worst performance measurements among word categories, and also the largest variation. One possibility of explaining it is the fact that these words may present particular challenges not only to subjects with some form of dysarthria, but also to normal participants, creating more blurried borders between classes. Furthermore, these are the only words for which subjects record a single take (compared to 3 takes for words in other categories), and so models are built out of training with a set 1/3 the size of the training sets for words in all other categories (digits, radio alphabet, etc.).

Figure 12 extends the results in Figure 11 of relative MinDTW and HausdorffDTW embeddings to the extended dataset (including 8 microphones per recording). Comparing MinDTW and HausdorffDTW in both the original and the extended datasets, it can be safely posited that HausdorffDTW produce worse classification results for the UA-Speech dataset. Being geared at reflecting a sort of worst-case distance between sets, or in our point-to-set case reflecting the distance to the extreme opposite border of the set of normal speech recordings of a word, Hausdorff enlarge the relative embeddings not only of the disturbed recordings, but of the normal recordings as well, which in the case of this data was reflected in a larger number of classification errors. That is not meant to discredit HausdorffDTW, whose utility in the general case is regarded as a function of how tightly close together are the in-class items in comparison to typical out-of-class items.

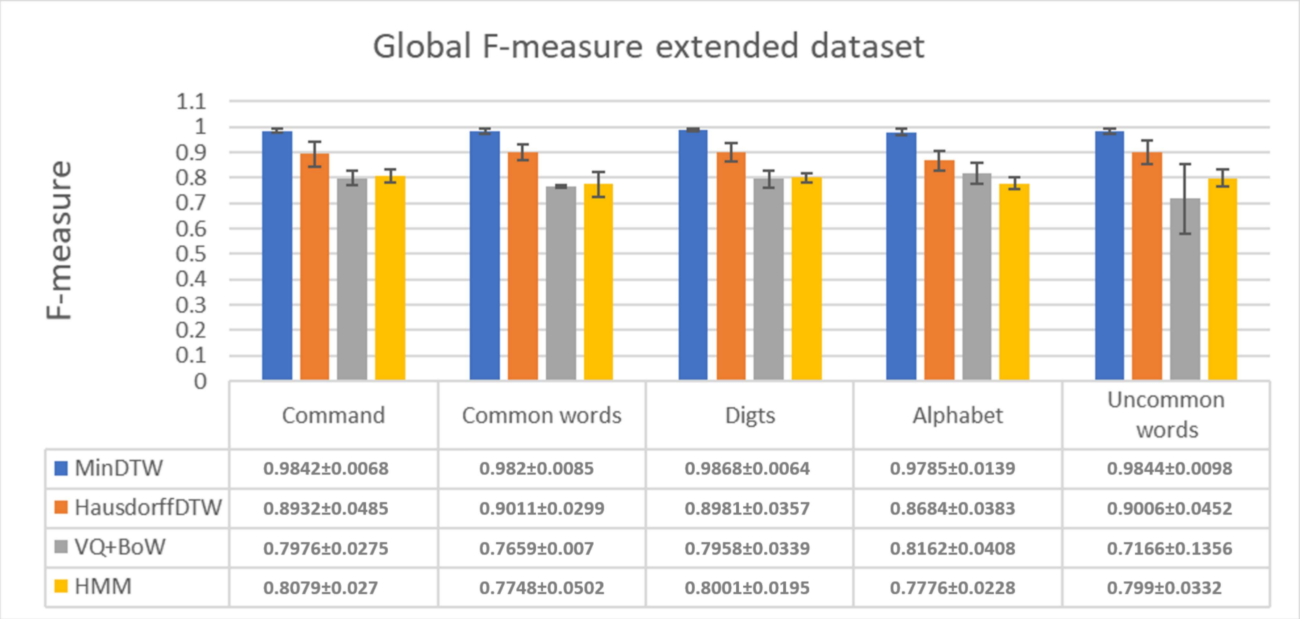
As for the comparison between the original and the extended datasets, it would be tempting to presume that the availability of more recordings of the same speech utterance would reinforce the classification mechanism and make it more robust to variations in timbre (arguably the main differing characteristic between the eight different but simultaneous microphone takes). This abductive interpretation is compatible with the improved performance observed for the MinDTW distance in Figure 12 compared to Figure 11. The fact that HausdorffDTW had the opposite outcome suggests that a more involved argument might clarify these empirical findings.

One difficulty encountered is the fact that the UA-Speech dataset labels subjects, rather than recordings, as either normal of dysarthric. But dysarthria does not manifest itself in every single vocal emission of a dysarthric subject. This entails situations in which an otherwise normal utterance would be used as disturbed in the classification mechanism, with a corresponding increase in False Negatives (FN), because the boundary of the disturbed class would be pushed into the class of normal recordings, lowering



	Command	Common words	Digts	Alphabet	Uncommon words
■ MinDTW	0.9518 ± 0.022	0.9533 ± 0.0233	0.9648 ± 0.0203	0.9506 ± 0.0194	0.9225 ± 0.0414
■ HausdorffDTW	0.8793 ± 0.0507	0.8783 ± 0.0396	0.8722 ± 0.0391	0.8391 ± 0.0664	0.8896 ± 0.0913
■ VQ+BoW	0.7976 ± 0.0275	0.7659 ± 0.007	0.7958 ± 0.0339	0.8162 ± 0.0408	0.7166 ± 0.1356
■ HMM	0.8377 ± 0.0339	0.8287 ± 0.0373	0.8342 ± 0.0231	0.8031 ± 0.0286	0.8258 ± 0.0558

Figure 11. Global F-measures for the original dataset, for each method and word category.



	Command	Common words	Digts	Alphabet	Uncommon words
■ MinDTW	0.9842±0.0068	0.982±0.0085	0.9868±0.0064	0.9785±0.0139	0.9844±0.0098
■ HausdorffDTW	0.8932±0.0485	0.9011±0.0299	0.8981±0.0357	0.8684±0.0383	0.9006±0.0452
■ VQ+BoW	0.7976±0.0275	0.7659±0.007	0.7958±0.0339	0.8162±0.0408	0.7166±0.1356
■ HMM	0.8079±0.027	0.7748±0.0502	0.8001±0.0195	0.7776±0.0228	0.799±0.0332

Figure 12. Global F-measures for the extended dataset, for each method and word category.

Recall=$\frac{\text{TP}}{\text{TP+FN}}$ and consequently F-measure.

Another experimental difficulty relates to the lack of pre-processing, because a significant portion of the dataset is affected by noise, other simultaneous voices and variable lengths of silence. This pre-processing was not the focus of the experiment, so severely affected recordings were manually removed from the dataset, in order to avoid a significant increase in the radius of the normal classes and a corresponding increase in the number of false positives (FP), affecting Precision=$\frac{\text{TP}}{\text{TP+FP}}$ and consequently F-measure.

One aspect worth mentioning, relating to how each technique uses the original data, has to do with quantization, which is bound to introduce noise. Relative DTW embeddings, being computed directly from the frame-based matrix representations, does not require any quantization step before actual classification. Bag-of-Words only makes sense in a quantized setting, due to the very nature of how histograms are built; hence there is an intrinsic relationship between the choice of the quantization parameter K (used in Kmeans) and the performance of a classification strategy based on BoW. We adopted the widely-used silhouette coefficient for the choice of this parameter, but a brute-force search over other values of K might uncover a better alternative. HMM can be defined over the original data, using continuous models for the emission probabilities; this was the approach adopted here. Alternatively, frame-based data can be quantized in order to produce smaller and possibly computationally more efficient discrete models for the emission probabilities. Preliminary experiments suggested that success rates were higher using the original data, so this was the standard adopted in our comparative experiment.

5. CONCLUSION

In this paper we have introduced a novel strategy for binary classification of heterogeneous time-dependent data via relative DTW embeddings. Items are mapped into relative point-to-class distances, from which suitable thresholds for classification may be computed.

The proposed strategy has proved to produce very good results within a speech classification context when compared to usual alternatives for this type of classification, namely Vector quantization followed by Bag-of-Words and Hidden Markov Models. Typical F-measure values for classifying disturbed versus normal word utterances in the UA-Speech dataset were above 0.95 for relative MinDTW embedding, against 0.71–0.81 for VQ+BoW and 0.77–0.83 for HMM.

One idea that looks worthwhile pursuing in the future is the possibility of multi-dimensional relative DTW embedding, considering several distances simultaneously, e.g. a bidimensional embedding of items into a

$$(\text{MinDTW},\text{HausdorffDTW})$$

relative distance space. The motivation for this is the fact that different distances may respond differently to in-class and out-of-class items, as discussed in Section 2; these subtleties, when taken simultaneously into account, could result in improved performance for binary classification problems involving frame-based feature matrices.

Acknowledgments

The first author acknowledges the support of FAPESP grant 2018/09373-8 and CNPq grant 309645/2016-6.

6. REFERENCES

[1] A. Bala, A. Kumar, and N. Birla, "Voice command recognition system based on mfcc and dtw," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335–7342, 2010.

[2] M. Shaneh and A. Taheri, "Voice command recognition system based on mfcc and vq algorithms," *World Academy of Science, Engineering and Technology*, vol. 57, pp. 534–538, 2009.

[3] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the musart testbed," *Journal of the Association for Information Science and Technology*, vol. 58, no. 5, pp. 687–701, 2007.

[4] M. Müller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization." in *ISMIR*. Victoria. Canada: Citeseer, 2006, pp. 192–197.

[5] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978, pp. 43 – 49 Volume: 26, Issue: 1, Feb 1978.

[6] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, pp. 2105–2108.

[7] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[9] J.-J. Aucouturier, F. Pachet, and M. Sandler, "The way it sounds: timbre models for analysis and retrieval of music signals," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1028–1035, 2005.

[10] T. Cho, R. J. Weiss, and J. P. Bello, "Exploring common variations in state of the art chord recognition systems," in *Proceedings of the Sound and Music Computing Conference (SMC)*. Citeseer, 2010, pp. 1–8.

[11] A. F. Machado and M. Queiroz, "A flexible and modular crosslingual voice conversion system." in *Joint ICMC / SMC*, 2014, pp. 1312–1319.

[12] J. Serra, E. Gómez, and P. Herrera, "Audio cover song identification and similarity: background, approaches, evaluation, and beyond," in *Advances in Music Information Retrieval*. Springer, 2010, pp. 307–332.

[13] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, K. W. T. Huang, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of Interspeech*, Brisbane. Australia, 2008, pp. 1741–1744.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech and Signal Processing, (ASSP)*, 1980, pp. 357–366, Volume:28.

[15] G. Jhawar, P. Nagraj, and P. Mahalakshmi, "Speech disorder recognition using mfcc," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, April 2016, pp. 0246–0250.

[16] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," in *SIGKDD Explor. Newsl.*, 2010, pp. 49–57, Volume 12, Issue 1.