Phone Classification in Pseudo-Euclidean Vector Spaces

Alexander Gutkin and Simon King

Centre for Speech Technology Research (CSTR)
University of Edinburgh, www.cstr.ed.ac.uk
Alexander.Gutkin@ed.ac.uk

Abstract

Recently we have proposed a structural framework for modelling speech, which is based on patterns of phonological distinctive features, a linguistically well-motivated alternative to standard vector-space acoustic models like HMMs. This framework gives considerable representational freedom by working with features that have explicit linguistic interpretation, but at the expense of the ability to apply the wide range of analytical decision algorithms available in vector spaces, restricting oneself to more computationally expensive and less-developed symbolic metric tools. In this paper we show that a dissimilarity-based distance-preserving transition from the original structural representation to a corresponding pseudo-Euclidean vector space is possible. Promising results of phone classification experiments conducted on the TIMIT database are reported.

1. Introduction

Current automatic speech recognition systems usually use Hidden Markov models (HMMs) of phones; speech is modelled as a linear sequence of these phones which have no explicit internal structure beyond the linear topology of the HMMs. Many researchers are looking for alternative approaches [1]. Recently, we presented a classification framework based on a structural representation of speech [2] in which phones are modelled as string templates, making use of the underlying phonological feature structure. Such a symbolic representation is motivated by the fact that a symbolic space is well-suited for capturing and exploiting structural properties of speech which vector space-based approaches fail to capitalise on.

Structural representations like this, while offering a greater representational freedom than conventional vector-space approaches, have their shortcomings, including the lack of much of the analytical machinery available in vector spaces. There are some symbolic space counterparts of well-known techniques, such as k-nearest neighbours [3], but their computational complexity is increased by the absence of vector space properties. Such limitations motivated us to draw on a theory which unifies structural and vector-space approaches, on one hand providing the representational convenience of symbolic spaces and on the other allowing us to use vector space decision-theoretical tools. It is such a theory, proposed in [4], that we consider in this paper for the representation of complex speech data.

2. Structural Representation

The lowest (i.e. closest to the speech waveform) level of our framework which is symbolic consists of a set of *phonological distinctive features*. Phonological distinctive features are seen in most varieties of phonology as the fundamental units out of which phonemes are constructed. We use a subset of one of the most popular feature systems used to represent speech, five *multivalued features*: front-back, place of articulation, manner of articulation, roundness and voicing. Each of these features takes one of several possible values – for example, manner of articulation is one of: approximant, fricative, nasal, stop, vowel, silence.

The neural networks which were used to recover these features from speech (refer to [1] for full details) use a 1-of-N_j encoding on their output units, hence there are N_j real-valued outputs (ranging $0 \rightarrow 1$) for each feature (for manner of articulation $N_j = 6$). The total number of such values produced by the neural networks for each frame is $N = \sum_{j=1}^{5} N_j = 25$. When classifying unknown speech patterns the output activations take continuous values between 0 and 1, and the features change value asynchronously. We map these continuous activation values into symbols using simple quantization over N separate finite alphabets of equal size (quantization level) for each of the N values separately.

The speech is now represented by a sequence of vectors of symbols; this can be seen as a sequence of symbolic matrices, each identifying a phone in terms of its distinctive phonological features (which henceforth will be referred to as *streams*). In the current work we restrict ourselves to classification: the phone boundaries are known but phone identities are unknown. A phone realization p is thus represented as:

$$\begin{matrix} f_1^{t_p} & f_1^{t_p+1} & \dots & f_1^{t_p+k_p-1} \\ f_2^{t_p} & f_2^{t_p+1} & \dots & f_2^{t_p+k_p-1} \\ \dots & \dots & \dots & \dots \\ f_N^{t_p} & f_N^{t_p+1} & \dots & f_N^{t_p+k_p-1} \end{matrix} \rightarrow_t$$

where t_p is start time and k_p is duration. Our symbolic phone classification system is template based: it consists of a set

of templates learnt from the training data, one or more per phone class to be recognized. The templates may be actual tokens from the training data, or may be constructed; in either case, each template is represented by the structure described above.

This representation has a number of attractive features: it accounts for duration and contextual effects; aspects of co-articulation such as assimilation can be accounted for, since the features (which can change value anywhere within a given template) are represented explicitly and independently. This paper only shows that we can make a transition from this representation to a vector space; we do not yet take full advantage of the structural representation.

2.1. Pseudo-Metric Space

Once the structural representation is obtained, the next step is to define a dissimilarity measure between pairs of templates, or between a template and a token to be classified. A simplifying assumption made in this paper is that the streams are independent and have equal importance.

A *phonological pseudo-metric space*, corresponding to our structural representation, is a pair (P, D) where P is a finite set of all possible templates having N streams and $D\colon P \times P \rightarrow \mathbb{R}^+$ is a mapping of the Cartesian product $P \times P$ into the set of non-negative real numbers $\mathbb{R}^+$, such that $D = \sum_{i=1}^{N} d_i$, where d_i can be any chosen string dissimilarity measure, satisfying the conditions of reflexivity: $\forall x \in P \;\; D(x,x) = 0$ and symmetry: $\forall x, y \in P \;\; D(x,y) = D(y,x)$. If the triangle inequality condition is satisfied in addition to the above, the resulting space is a metric space. The resulting properties of the pseudo-metric space are dictated by the per-stream distance functions d_i (which are currently the same for all the streams). In the remaining discussion we refer to phonological feature templates as *patterns*.

3. Vector Representation

The pseudo-metric space (P, D) as defined above, replaces the notion of *similarity* of two patterns by the peer notion of *dissimilarity* expressed by the distance function, which allows us to reduce the original pattern to a point in some abstract vector space where decisions (e.g. classifications) are to be made based on the metric information only.

The symmetric dissimilarity matrix between the patterns of the set P needs to be preserved by the embedding of the symbolic space into a vector space. It has been shown that, given a pseudo-metric space, it is always possible to construct an isometric mapping onto the corresponding pseudo-Euclidean vector space (section 3) – a member of a class of spaces in which symmetric bilinear forms are not restricted to be positive – and that in many cases such a construction cannot be accomplished in a classical Euclidean space [4].

3.1. Pseudo-Euclidean Spaces

A *pseudo-Euclidean* space $\mathbb{R}^{(n_+,n_-)}$ [4] of *signature* (n_+, n_-) – $n_+, n_- \geq 0$ – is a pair (V, Φ) where V is a real vector space of dimension $n = n_+ + n_-$ and Φ is a non-degenerate symmetric bilinear form of signature (n_+, n_-), which measures the inner product in V. Given an orthonormal (w.r.t Φ) basis $(e_i)_{i\in[1,n]}$, the inner product between the two vectors $x, y \in V$ is given by $\langle x, y\rangle = \sum_{i=1}^{n_+} x^i y^i - \sum_{j=n_++1}^{n} x^j y^j$. The above space can be viewed as consisting of two non-commensurable Euclidean subspaces of dimensions n_+ and n_-, respectively. If $n_- = 0$, the pseudo-Euclidean space corresponds to a Euclidean space. The *square of the distance* (which can be negative) between the two vectors in pseudo-Euclidean space is defined as $\|x-y\|^2 = \langle x-y, x-y\rangle = (x-y)^T J(\Phi)(x-y)$, where $J(\Phi) = \left(\begin{smallmatrix} I_{n_+\times n_+} & 0 \\ 0 & -I_{n_-\times n_-} \end{smallmatrix}\right)$ is the canonical matrix of the symmetric bilinear form corresponding to the orthonormal w.r.t. Φ basis $(e_i)_{i\in[1,n]}$ of V and I denotes an identity matrix.

3.2. Linear Embedding

Given a finite pseudo-metric space (P, D), $P = \{p_i\}_{i=1}^k$, there exists [4] an *isometric embedding* $\alpha\colon (P, D) \rightarrow \mathbb{R}^{(n_+,n_-)}$. In other words, let $v_i = \alpha(p_i)$, $i \in [1, k]$, then for all patterns p_i and p_j in the original set P, $\|v_i - v_j\| = D(p_i, p_j)$.

An algorithm which constructs the vector representation assumes, without loss of generality, that the mean vector of the vector representation α coincides with the origin, i.e. $\overline{v} = \frac{1}{k}\sum_{i=1}^{k} v_i = 0$. In this case, it can be shown that the non-zero characteristic values of the $k\times k$ matrix $M(\Phi) = (m_{i,j})$ of the symmetric bilinear form Φ of (P, D), where

$$m_{i,j} = \frac{1}{2}\Big[\frac{1}{k}\Big(\sum_{i=1}^{k} D(p_i,p_j)^2 + \sum_{j=1}^{k} D(p_i,p_j)^2\Big) - \frac{1}{k^2}\sum_{i,j=1}^{k} D(p_i,p_j)^2 - D(p_i,p_j)^2\Big] ,$$

coincide with those of the covariance matrix of (P, D) w.r.t α [4]. The vector representation is thus constructed by computing $M(\Phi)$ and performing its eigen-decomposition obtaining $M(\Phi) = EFE^T$, where E is the matrix of the eigenvectors and F is a diagonal matrix of the eigenvalues. By reorganizing F into another diagonal matrix C containing first the positive eigenvalues of $M(\Phi)$ in decreasing order, then the magnitudes of the negative eigenvalues in decreasing order followed by zeros, one obtains

$$M(\Phi) = HCH^T = HC^{\frac{1}{2}}\left(\begin{smallmatrix} J & \\ & 0 \end{smallmatrix}\right)C^{\frac{1}{2}}H^T = U\left(\begin{smallmatrix} J & \\ & 0 \end{smallmatrix}\right)U^T ,$$

where H is the matrix of the eigenvectors corresponding to the eigenvalues of $M(\Phi)$ in C and $J_{n\times n}$ is a canonical matrix of Φ from section 3.1. The first $n_+ + n_-$ elements of the i-th row of U, where $U = HC^{\frac{1}{2}}$, define the coordinates of $\alpha(p_i)$, $i \in [1, k]$, of a vector representation $\alpha\colon (P, D) \rightarrow \mathbb{R}^{(n_+,n_-)}$ w.r.t. an orthonormal basis of $\mathbb{R}^{(n_+,n_-)}$. The number of negligible eigenvalues (corresponding to noisy dimensions) of C is usually small, hence n is usually close to k.

3.3. Dimensionality Reduction

Since the eigenvalues $M(\Phi)$ correspond to the characteristic values of the generalized covariance matrix of the set $\{\alpha(p_i)\}$, the *reduced* vector representation $\beta\colon (P,D) \rightarrow \mathbb{R}^{(m_+,m_-)}$, where $m = m_+ + m_- < n$, can be constructed from α by the mapping $\gamma\colon \mathbb{R}^{(n_+,n_-)} \rightarrow \mathbb{R}^{(m_+,m_-)}$, which is an orthogonal projection of the exact representation α on the subspace spanned by the corresponding principal axes of the covariance matrix [4]. This is accomplished by removing the axes corresponding to small magnitudes of the eigenvalues $|c_i|$ of C and retaining the eigenvalues corresponding to principal uncorrelated axes. If the removed eigenvalues are small, the resulting configuration $\beta = \gamma \circ \alpha$ possesses the same isometric properties as α.

3.4. Metric Projection of Unseen Patterns

During the classification stage, an orthogonal projection of the new pattern p onto $\mathbb{R}^{(m_+,m_-)}$ is found by assuming that p maps to a point in $\mathbb{R}^{(n_+,n_-)}$ with the calculated distances to k vectors $\beta(p_i)$ in $\mathbb{R}^{(m_+,m_-)}$. Such a construction allows us to avoid reembedding anew every time an unseen pattern is presented [4]. The construction begins by performing a parallel translation $\tau\colon \mathbb{R}^{(m_+,m_-)} \rightarrow \mathbb{R}^{(m_+,m_-)}$, $\tau(v_i)_{1\leq i\leq k} = v_i - v_0$, where $v_i = \beta(p_i)$ are the vectors comprising the representation and $v_0 = \beta(p_0)$ is a fixed representation of an origin, chosen from the set P as a pattern whose average distance to the rest of the patterns in the training set is minimum.

Let $u_j = \tau(v_j)_{j\in[1,m]}$, be the chosen basis of $\mathbb{R}^{(m_+,m_-)}$ whose $m \times m$ Gram matrix $G = (\langle u_i, u_j\rangle)_{i,j}$ has signature (m_+, m_-). Metric projection $\delta\colon (P,D) \rightarrow \mathbb{R}^{(m_+,m_-)}$ of new patterns onto $\mathbb{R}^{(m_+,m_-)}$ is specified so that unique projection of new pattern p is defined by $m+1$ distances $D(p,p_0)$, $D(p,p_i)_{i\in[1,m]}$ as $\delta(p) = UG^{-1}b$, where columns of U are the coordinate columns of m vectors u_i and $b_{m\times 1}$ is a vector whose ith coordinate is given by $\frac{1}{2}\left[D(p,p_0)^2 + D(p_i,p_0)^2 - D(p,p_i)^2\right]$ [4]. Since $B = UG^{-1}$ can be precomputed during the training stage, the only online computations involved are those of b and the product Bb.

Since the reduced vector representation gives an approximation of the original finite metric set, the Gram matrix G for $\delta(p_i)$ in $\mathbb{R}^{(m_+,m_-)}$ differs from the exact Gram matrix for the $\alpha(p_i)$ in $\mathbb{R}^{(n_+,n_-)}$, while the calculation of vector b is based on the precise distances. In order to avoid this perturbation, an alternative construction called *corrected metric projection*, referred to as δ_C, is suggested in [5]. The correction is achieved by projecting the points $\alpha(p_i)$ onto the subspace of $\mathbb{R}^{(n_+,n_-)}$ spanned by a subset of the vectors close to the reduced space $\mathbb{R}^{(m_+,m_-)}$ and then, projecting them back onto $\mathbb{R}^{(m_+,m_-)}$.

For both metric projection methods, δ and δ_C, the m basis vectors spanning $\mathbb{R}^{(m_+,m_-)}$ are chosen in such a way as to minimize the average projection error between the projection of the entire training set (obtained with δ or δ_C) and the original vector representation obtained by linear embedding (α or β) of a pseudo-metric space [5, 6].

4. Experiments

In this section we present some of the experimental results of a phone classification task on the data represented in the pseudo-Euclidean domain. The experiments use the TIMIT database [7]. This is a corpus of high-quality recordings of read continuous speech from North American speakers. The entire corpus is reliably transcribed at the word and surface phonetic levels. For details of the feature-detecting neural networks mentioned in section 2, please refer to [1].

The standard training/test data partition is kept, with only the `sx` and `si` sentences being used (3696 training utterances from 462 different speakers, less 100 sentences held out for cross-validation training of neural networks; 1344 test utterances from 168 speakers). No test speakers are in the training set; there are 39 phone classes. Based on classification results obtained on the structural representation [2], we chose a quantization level of 10 and the metric D was a weighted Levenshtein edit distance.

4.1. Three-class Problem

The first experiment uses three phones which are *a priori* known to be reasonably separable: `aw` (low back round vowel) `b` (voiced bilabial stop) and `z` (voiced alveolar fricative). The original training set contains 6629 patterns. Since the matrix of pseudo-metric space interdistances for this set is rather large for matrix decomposition algorithms, we used clustering [2] to obtain a smaller training set of 100 patterns per class. The entire test set (2423 patterns) was used. The first two plots of figure 1 show the corrected metric projection δ_C (visualized w.r.t. the three principal axes) of the data onto the vector space representation constructed using linear embedding α (section 3). The visualization of the three principal axes of the corrected metric projection suggests that linear decision surfaces can separate the classes; we used a feed-forward neural network with activation units mapped directly to target units (denoted LDS). We compare this to a vector-space equivalent of $k-$NN AESA [3] (denoted kNN).

Given the finite pseudo-metric space (P,D) and dimension m, two different reduced vector representations were constructed using regular δ (denoted by subscript R) mapping and corrected δ_C (subscript C) mappings. Two different approaches to basis selection of the reduced vector space $\mathbb{R}^{(m_+,m_-)}$ were employed: the *regular* approach (superscript R) [6] ignores the class label of vectors, which can result in uneven representation of classes within the basis of the reduced space; a novel *class*-based approach (superscript C) selects the basis of the reduced space to minimise projection error whilst ensuring the basis vectors are well-balanced in terms of class representation. Best classification results (chosen from reduced dimensions of $m < 150$) are shown in table 1. For kNN classifiers, $k = 1$ outperformed $k > 1$.

As can be seen from table 1, LDS consistently outper-

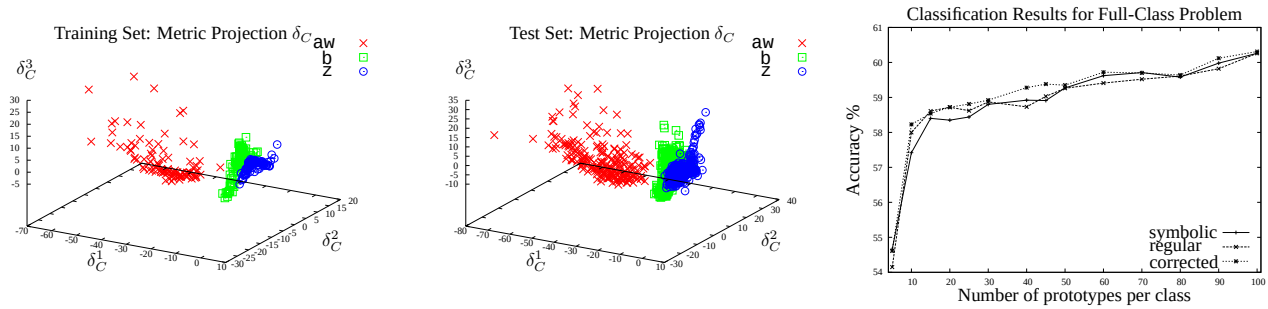

Figure 1: Corrected metric projections δ_C for the three-class problem and full-problem classification results.

Method	m	Error (%)
LDS_R^R	39	0.7
LDS_C^R	144	0.6
kNN_R^R	85	2.8
kNN_C^R	63	1.1

Method	m	Error (%)
LDS_R^C	33	0.6
LDS_C^C	130	0.5
kNN_R^C	73	2.4
kNN_C^C	50	1.0

Table 1: Best classification error rates for the three class problem. Best result obtained in the pseudo-metric space with 1-NN AESA was 0.9%.

forms k-NN in both symbolic and vector spaces. The class-based basis selection improves the performance of all models. The 1-NN classifier in the *reduced* pseudo-Euclidean space does not seem to handle perturbations introduced by the dimensionality reduction as well as the neural networks. Its error rate, however, comes close to its symbolic counterpart by *only* using around 17% (50 out of 300 patterns) of the original training data.

4.2. Full Problem

The full 39-class task consists of 124962 training patterns and 46633 test patterns. The training set was pre-clustered to produce much smaller training sets with different numbers of prototypes (from 5 up to 100) per class [2]. For each of these training sets, two different reduced vector representations were constructed using regular δ and corrected δ_C projections. For all the training sets, and for dimensions m higher than 10, the resulting vector spaces are pseudo-Euclidean. In order to preserve the significant axes of the representation, only the relatively small eigenvalues (less than 10^{-4}) were removed.

Results are shown in figure 1; optimal $k = 1$ [2]. For both regular and corrected constructions, k-NN AESA in pseudo-Euclidean space appears to consistently outperform the pseudo-metric counterpart for all the dimensions corresponding to number of prototypes per class of up to 50. The best results were obtained for 100 prototypes per class: 60.26% for regular metric projection and 60.31% for the corrected one, compared to the best result of 60.26% reported [2] for the pseudo-metric space.

5. Conclusions and Future Work

We have demonstrated the feasibility of a transition from symbolic to vector space by the construction of pseudo-Euclidean embeddings of finite pseudo-metric spaces. The vector-space classification accuracy is very close to that in the original symbolic space, confirming previous findings [4, 6]. Since the future directions of our research will include more structurally complex representations of speech, reliable construction of robust dissimilarity-based pseudo-Euclidean space representations of pseudo-metric spaces is of paramount importance. In addition, we are planning to use non-linear classifiers, which are available for pseudo-Euclidean vector spaces [8]).

6. References

[1] M. Wester, "Syllable classification using articulatory-acoustic features," in *Proc. Eurospeech*, Geneva, 2003, pp. 233–236.

[2] A. Gutkin and S. King, "Structural Representation of Speech for Phonetic Classification," in *Proc. 17th International Conference on Pattern Recognition*, vol. 3, Cambridge, UK, Aug. 2004, pp. 438–441.

[3] A. Juan and E. Vidal, "On the Use of Normalized Edit Distances and an Efficient k-NN Search Technique (k-AESA) for Fast and Accurate String Classification," in *Proc. ICPR*, vol. 2, Sept. 2000, pp. 680–683.

[4] L. Goldfarb, "A New Approach to Pattern Recognition," in *Progress in Pattern Recognition*, L. N. Kanal and A. Rosenfeld, Eds. Amsterdam: Elsevier, 1985, vol. 2, pp. 241–402.

[5] ——, "Metric Data Models and Associative Memories," in *Proc. IASTED RAI/IPAR*, vol. 3, Toulouse, France, June 1986, pp. 53–73.

[6] E. Pękalska and R. P. Duin, "Prototype selection for finding efficient representations of dissimilarity data," in *Proc. ICPR*, vol. III, 2002, pp. 37–40.

[7] J. S. Garofolo, *Getting Started with the DARPA TIMIT CD-ROM: an Acoustic Phonetic Continuous Speech Database*, NIST, Gaithersburgh, Maryland, 1988.

[8] E. Pękalska, P. Paclík, and R. P. Duin, "A Generalised Kernel Approach to Dissimilarity Based Classification," *JMLR*, vol. 2, no. 2, pp. 175–211, 2002.