EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Broad phonetic class definition driven by phone confusions

Carla Lopes[1,2*] and Fernando Perdigão[1,3]

Abstract

Intermediate representations between the speech signal and phones may be used to improve discrimination among phones that are often confused. These representations are usually found according to broad phonetic classes, which are defined by a phonetician. This article proposes an alternative data-driven method to generate these classes. Phone confusion information from the analysis of the output of a phone recognition system is used to find clusters at high risk of mutual confusion. A metric is defined to compute the distance between phones. The results, using TIMIT data, show that the proposed confusion-driven phone clustering method is an attractive alternative to the approaches based on human knowledge. A hierarchical classification structure to improve phone recognition is also proposed using a discriminative weight training method. Experiments show improvements in phone recognition on the TIMIT database compared to a baseline system.

Introduction

Broad phonetic classes (BPC) have widely been used in speech recognition research as, for instance, automatic language identification [1]; speaking rate estimation [2]; multi lingual systems [3,4] and, especially, in phone recognition [5-8]. Its success is due to the carried additional information that contributes to improve the recognition rates, especially under noise conditions [9]. In phone recognition task, BPC information may be used as an additional set of acoustic features or it may be integrated in the phone predictions. One successful example is presented by Siniscalchi et al. [6], one of the best results reported on the phone TIMIT recognition task, where 15 broad articulatory classes are used to predict posterior phone probabilities and to rescore phone lattices. Another interesting example is given by Morris and Fosler-Lussier [7] where the outputs of eight broad class classifiers are used as input features on a conditional random field (CRF) model. In [8], a hierarchical classification from successive broad classes is used with improvements in phone accuracy. Several other studies related to BPC could be referred, e.g. [5,10], sometimes using different terminology. In literature, BPC take different names (such as broad phonetic groups, speech attributes, events, etc.) but in all these studies they are always sets of phones with similar acoustic/phonetic features drawn manually by an expert (knowledge-driven information). The broad classes are selected according to acoustic-phonetic properties that derive from articulatory constraints or from hearing perception. This selection may contain some subjectivity or may be difficult to carry out when dealing with other kind of speech units like syllables or when considering coarticulation phenomena.

BPC for the English language is an issue that has widely been addressed by the scientific community [5,7,8,11]. Its definition is usually related to the manner and place of articulation, so that all the broad classes show good agreement within some phonetic, articulatory and/or acoustic properties. The efficient construction of smaller/more compact phone sets is not trivial, however, as is confirmed by the lack of consensus between several proposals. For example in the TIMIT corpus, in [12], phone [dx] is classified as a sonorant consonant, while in [13] the same phone is classified as a stop and Halberstadt and Glass [10] classify it as a nasal/flap. The same happens with phone [hv]. In [12], it is a sonorant consonant and in [5,10] it is a fricative. Comparing the "fricative" broad classes of [5,14] it can be seen that the first proposal includes the phone [hv] while the second does not. The above examples are to show the subjective nature of the approaches based on human knowledge, even when referring to such a widely studied language as American English.

* Correspondence: calopes@co.it.pt
[1]Instituto de Telecomunicações, Polo II, Coimbra P-3030-290, Portugal
[2]ESTG, Instituto Politécnico de Leiria, Campus 2, Leiria P-2411-901, Portugal
Full list of author information is available at the end of the article

This subjectivity affects not only the number of classes, but also the set of phones in each category. These problems, related to the expert-driven approaches, have prompted the emergence of data-driven approaches, where the classes' composition is guided by data. We explore in this article a data-driven method where the broad classes are automatically defined according to the output of a speech recognition system. The method can be applied in systems using all kinds of recognition units (phones, syllables, subwords, phones of different languages, etc.) and do not conduct to a static division.

The output of a classification system is usually evaluated comparing all the recognized sequences with the corresponding references. From this comparison, a confusion matrix can be computed. In the proposed approach it is considered that if a unit (phone or other) has much confusion with other unit is because, to the recognizer, they are somehow similar. This approach may not fully agree with phonetics principles or acoustic theories but it can be very helpful allowing to overcome deadlocks in some situations. Take the case of a multilingual system, where a knowledge-based approach has to involve experts from all the languages involved. The same occurs when dealing with varieties of a language (e.g., European Portuguese or Brazilian Portuguese). The proposed automatic clustering method extends the possibility of using broad classes' information to systems based on other than the phone unit.

Data-driven clustering usually stands for a statistical measurement of a class. The key point of all clustering algorithms is the choice of a proximity or distance measure. This measure can be obtained from acoustic models, e.g., [3,15], or even rely on the confusion matrix, e.g., [4,16]. Model-driven methods and confusion-driven[a] methods are then the two major categories of data-driven phone clustering algorithms.

In model-driven methods, the acoustic similarity between two phones can be achieved from the theoretical distance between the corresponding acoustic models. This distance can be the Bhattacharyya distance between two Gaussian mixture models [3,15]; a relative entropy-based (Kullback–Leibler divergence) distance between two Laplacian mixtures [17], etc. In [18], a data-driven phonetic broad class generation is proposed where mutual information is used to compute similarity between models, while in [19] a similarity measure based on the likelihood between the acoustic frames and the hidden Markov models (HMMs) is proposed.

In this article, we propose a confusion-driven method to generate phone clusters. This approach was already proposed in [4], where the phones are grouped using rules depending on a set of weights and thresholds. Our proposal differs from [4] in so far as we define a metric to compute phone distances.

The remainder of the article is organized as follows. In Section 2, we introduce the data-driven approach and Section 3 presents the metric defined that evaluates the phones similarities. It also describes the way that the confusion matrix (which is the base for phone similarities computation) was achieved. Section 4 presents a system where broad classes are used in order to enhance phone recognition, and in Section 5 experimental results, comparing data-driven and knowledge-driven approaches, are presented. Finally, some conclusions and future improvements are drawn in Section 6.

Data-driven broad classes

In confusion-driven methods, the similarity measure approach comes from the phone confusion matrix M. This matrix is computed from the output of a phone recognizer by aligning the recognized sentence with the reference one, using a dynamic programming algorithm (the *Levenshtein* algorithm). It includes the concept of hits and confusions, as well as insertions (INS) and deletions (DEL). An example of part of a confusion matrix of our phone recognizer using TIMIT [20] data is shown in Figure 1. The diagonal refers to the number of correctly recognized phones (hits) and the off-diagonal elements refer to the number of misclassifications.

The idea behind confusion-driven methods is that similar phones tend to be more confusable and should belong to the same class. Figure 2a shows, in pseudo colours, a confusion matrix using the 61 original TIMIT phones, where the phones are in alphabetical order and with blue representing the lower value and dark red the higher value. If highly confusable phones are repositioned in such a way that they become near each other, we get the matrix depicted in Figure 2b. In this second matrix, we can easily distinguish several sets (clusters) of phones where confusion between all the elements of the cluster is much higher than between other phones or phone clusters. In an attempt to find this set of clusters, this study explores

		RECOGNISED AS						
		aa	ae	ah	ao	aw	ax	DEL
SOURCE REFERENCE	aa	456	8	52	87	16	3	125
	ae	12	448	23	2	10	5	88
	ah	45	31	369	16	9	71	111
	ao	67	2	21	441	8	7	113
	aw	16	14	6	6	121	0	14
	ax	5	2	64	17	8	592	217
	INS	21	21	19	17	14	30	

Figure 1 Part of a confusion matrix: output of our TIMIT-based phone recognition system.

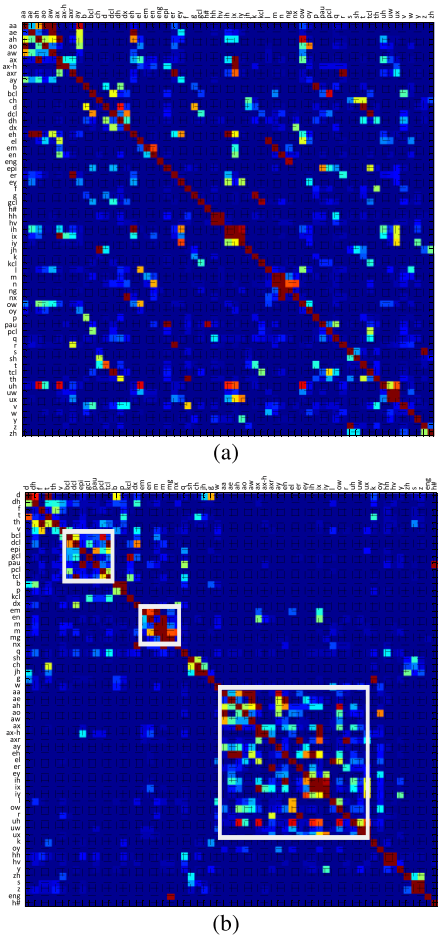


Figure 2 (a) Confusion matrix of our TIMIT-based phone recognition system. (b) A repositioned version of the matrix depicted in **(a)** where highly confusable phones are place near each other.

an automatic classification method, where the phonetic classes are found according to the output of an automatic phone recognition system. Phones are grouped according to a similarity measure estimated from the confusion matrix given by the recognizer performance.

Confusion-driven phone distance measure

This section describes an automatic method for phonetic class generation based on a confusion matrix of a phone recognizer. Usually, the clustering techniques involve three concepts:

1. A data model;
2. A proximity criterion (similarity, distance, etc.);
3. A clustering algorithm that generates the clusters (broad classes) using the data model and the similarity measure.

These concepts are discussed in following sections.

Data model

In the proposed method, the data comes from a confusion matrix yielded by the performance analysis of a phone recognition system. This matrix may contain results at the frame level, if artificial neural network (ANN), support vector machines (SVM), or CRF-based recognizers are used, or at the segment level, if segment models are used, such as HMMs or hybrid systems, as in the present case. An MLP/HMM hybrid system is used that combines an overall HMM structure with the class predictions given by a multilayer perceptron (MLP) classifier, thus benefiting from the time modelling abilities of HMMs and the discrimination capabilities of ANNs. In the TIMIT training set, this involves the recognition of 143 k segments.

Hybrid MLP/HMM description

An MLP network, with a single hidden layer, was trained for phone classification at a frame level. The last layer performs a 1-to-61 classification over the set of (TIMIT) phones. Speech was analysed every 10 ms with a 25-ms Hamming window. Thirty-nine parameters were used as standard input features representing 12 Mel Frequency Cepstral Coefficients (MFCCs), plus energy, and their first and second time-derivative coefficients. The context window used was 170 ms but only 9 frame features were used, as described in [8]. The current frame is in the centre of the context window (temporal information of past and future is included). The *softmax* function was used as the activation function of the output layer so that the output values could be interpreted as posterior probabilities. The hidden layer has 1,000 nodes and uses a sigmoid activation function. All the network weights and bias are adjusted using batch training with the resilient back-propagation (RPROP) algorithm [21], so as to reduce the minimum-cross-entropy error between network output and the target values. The network has 413 k free training parameters.

HMMs were built beforehand for each phone using HTK3.41, [22], in order to estimate the transition probabilities between states. Each phone was modelled by a three-state left-to-right HMM and each state was modelled by a single Gaussian model. The input features were the same as for the MLP.

In the hybrid MLP/HMM system, the state likelihoods are replaced by the posterior probabilities given by the output predictions of the MLP. The three states share the same MLP output. We used HTK [22], with some changes in order to replace the usual Gaussian mixture models with the outputs of the MLP.

The confusion matrix (data model) is computed using the entire TIMIT training set, which consists of all si and sx sentences of the original training set (3,698 utterances). The performance of the hybrid system is usually evaluated by means of Correctness (Corr) and Accuracy (Acc). We have used HTK evaluation tool HResults to

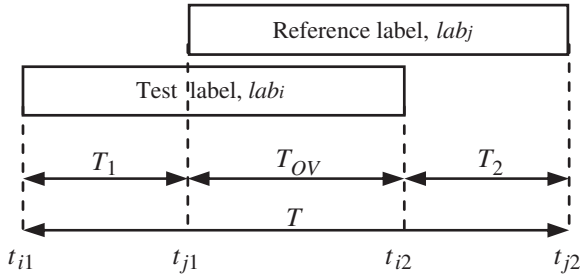


Figure 3 Measurement of the time misalignment between two labels.

compute them. But with this tool only the token sequence is taken for evaluation. A fine evaluation takes into account not only the correct identified sequence of phones, but also their time localization. This was the criteria used in the confusion matrix generation on which the proposed clustering is based. A brief description of the evaluation procedure used is given below. A full description can be found in [23].

The evaluation procedure uses a modified *Levenshtein* algorithm [24] in the alignment between labelled and recognized phones, where the degree of overlapping between them is taken in the local distance definition. This algorithm finds the best alignment between two strings and inserts a penalty if an error occurs (insertion, deletion and substitution), but no penalty is applied if the labels match. In our proposal, we include an additional penalty that is proportional to the average of the left and right misalignments. If the labels do not overlap ($T_{\mathrm{OV}} \leq 0$ in Figure 3), this penalty is set to a maximum value ($p_{\max}$), such that an insertion or a deletion will be preferred to a misaligned substitution.

Taking t_{i1},t_{i2} and t_{j1},t_{j2} as the boundaries of the test and reference labels, as indicated in Figure 3, they overlap if $t_{i2} > t_{j1}$ or $t_{j2} > t_{i1}$ and then the total time of the labels is

$$\begin{aligned} T &= \max(t_{i2}, t_{j2}) - \min(t_{i1}, t_{j1}) \\ &= T_1 + T_2 + T_{OV} \end{aligned} \tag{1}$$

and the overlapping time is

$$T_{\mathrm{OV}} = \min(t_{i2}, t_{j2}) - \max(t_{i1}, t_{j1}) \tag{2}$$

The left and right misalignments are $T_1 = |t_{j1} - t_{i1}|$ and $T_2 = |t_{j2} - t_{i2}|$. If the labels lab_i and lab_j match their name but are not perfectly aligned, then we introduce an additional association penalty, $p_A(i,j)$, which is inversely proportional to the overlap between the labels, according to the following expression:

$$p_A(i,j) = \frac{(T_1 + T_2)/2}{T_{\mathrm{OV}}} = \frac{1}{2}\left(\frac{T}{T_{\mathrm{OV}}} - 1\right) \tag{3}$$

If the labels overlap more that 50%, p_A is smaller than 0.5. As far as the overlapping decreases, this distance increases and is clipped to $p_{\max} = 15$, which corresponds to 3.2% of overlapping. In order to promote confusions (substitutions) we penalize insertions and deletions significantly by setting $p_{\mathrm{INS}} = 12$ and $p_{\mathrm{DEL}} = 12$. The optimal alignment is found by tracing back the path of accumulated penalties from the last label pair to the origin of the (i, j) grid. With the alignment of all labels a confusion matrix can then be computed. This matrix represents the data model referred to in the beginning of Section 3.

Similarity measure

The confusion matrix is often converted into a symmetric similarity matrix using the so-called *Houtgast* algorithm. It measures the similarity between reference phones i and j using the number of confusions of these phones with all other phones k. More specifically, if N is the total number of classes (phones), the *Houtgast* similarity between phones i and j, s_{ij} is given by

$$s_{ij} = s_{ji} = \sum_{k=1}^{N} \min(f_{ik}, f_{jk}) \tag{4}$$

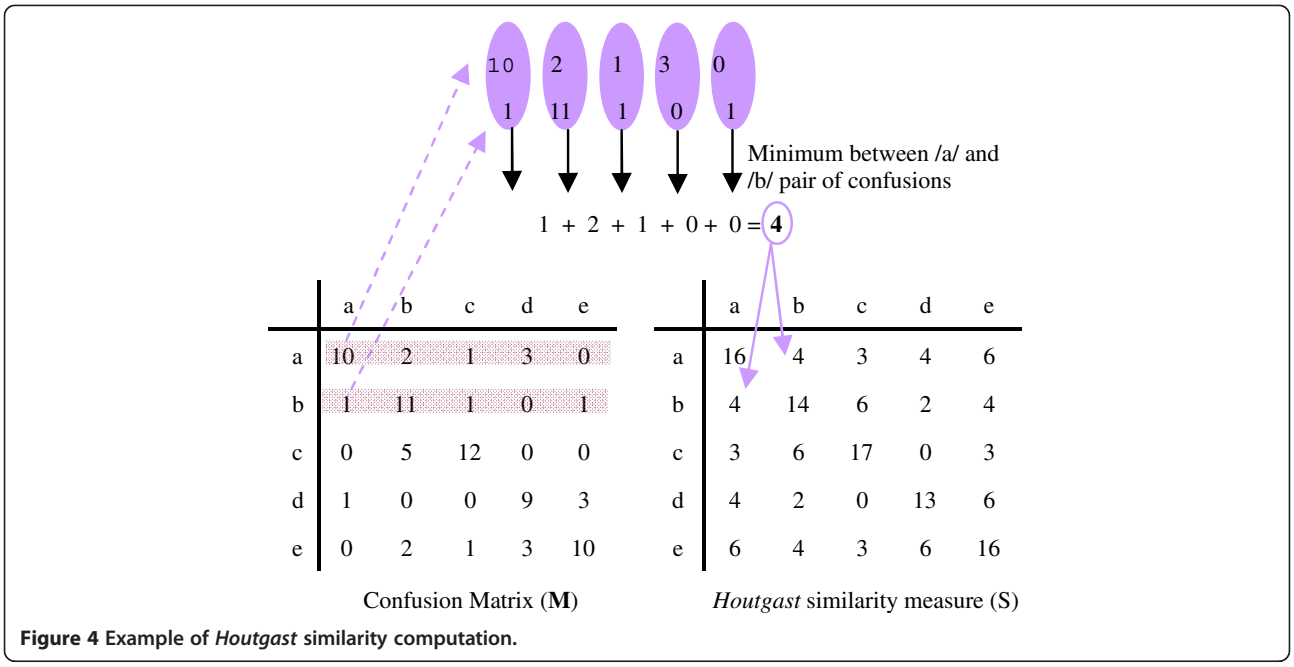
with $1 \leq i, j \leq N$. If $i = j$,f_{ij} is the number of hits instead of confusions. According to this measure, two phones i and j are similar if they both have many confusions with the same phones and their similarity is zero only when phones i and j have no simultaneous confusions with any phones. Figure 4 gives a simple example of this measure.

The confusion matrix of a good phone recognizer is close to a diagonal matrix. The out-diagonal values represent misclassified phones (confusions between phones). Since the numbers of occurrences of each phone are quite different (due to phonetically unbalanced speech material) we normalize the confusion matrix by dividing the frequency counts f_{ij} by the total number of occurrences of the phone i in the speech database:

$$p_{ij} = \frac{f_{ij}}{\sum_{n=1}^{N} f_{in}} = P(\hat{c}_j | c_i) \tag{5}$$

In this way, we have an estimate of the probability of recognizing the model or cluster $\hat{c}_j$ when its reference class is c_i. In this case the *Houtgast* similarity measure becomes

$$\begin{aligned} s'_{ij} &= \sum_{n=1}^{N} \min(p_{in}, p_{jn}) \\ &= \sum_{n=1}^{N} \min(P(\hat{c}_i | c_n), P(\hat{c}_j | c_n)) \end{aligned} \tag{6}$$



	a	b	c	d	e
a	10	2	1	3	0
b	1	11	1	0	1
c	0	5	12	0	0
d	1	0	0	9	3
e	0	2	1	3	10

	a	b	c	d	e
a	16	4	3	4	6
b	4	14	6	2	4
c	3	6	17	0	3
d	4	2	0	13	6
e	6	4	3	6	16

Figure 4 Example of *Houtgast* similarity computation.

This new measure has the following properties: (i) $s'_{ij} \leq 1$ and (ii) $s'_{ii} = 1$. Because $\min(a,b) = \frac{1}{2}(a + b - |a - b|)$ and because $\sum_{n=1}^{N} p_{in} = 1$, a distance measure between phones i and j can therefore be defined as

$$d_1(c_i, c_j) = 2\left(1 - s'_{ij}\right) = \sum_{n=1}^{N} |p_{in} - p_{jn}| \tag{7}$$

This distance forms a metric because it is the L_1 norm applied to row differences of matrix P (with elements p_{ij} as defined in (5)). Several similarity measure proposals can be found in literature [4,16,25], but as referred in [26], they do not fulfil the properties of a proper metric. In the present proposal, d_1 has the three metric properties: it is positive, symmetric and satisfies the triangle inequality.

Other distance measures can be defined based on the same principle, in particular the Euclidean distance of rows,

$$d_2(c_i, c_j) = \sqrt{\sum_{n=1}^{N} (p_{in} - p_{jn})^2} \tag{8}$$

Clustering method

The proposed clustering method groups phones in a multilevel hierarchy where clusters at one level are combined as clusters at the next level. Clustering can be achieved following hierarchical agglomerative clustering paradigm [27,28] as follows. Initially, each phone will be considered as a distinct cluster.

Step 1 → *Compute matrix P using (4).*
Step 2 → *Find the distance between each pair of phones using* (7).
Step 3 → *Compute the distances between all clusters. The distance between clusters r and s can be computed with several criteria, of which the simplest is the nearest neighbour:* $d(r,s) = \min\left(d\left(c_{i|r}, c_{j|s}\right)\right), i = 1 \ldots n_r; j = 1..n_s$ *where* n_k *is the number of phones in cluster k and* $c_{i|k}$ *is the ith phone in cluster k.*
Step4 → *Create a new cluster by grouping the two nearest ones.*

Steps 3 and 4 will repeat until the number of desired clusters or a distance threshold is reached. Another possibility is to compute clusters until all the phones belong to the same cluster. A dendrogram can be built from these clusters, to allow the decision, the level or scale of clustering that is most appropriate for the application.

TIMIT phone clustering results

Starting from the data model described in Section 3.1 and following the steps proposed in Section 3.3, we arrive at a hierarchical phone clustering, depicted as a binary cluster tree (dendrogram) in Figure 5. The labels along the horizontal axis represent the phones in the original data set and vertical axis refers to the distance between the phones. This distance is computed from a similarity measure using $d_1(c_i, c_j)$. The number of clusters depends on where the tree is cut. In the example given in Figure 5, the 61 TIMIT phones are represented by the 9 clusters presented in Table 1.

The consistence of the clusters, particularly the ones involving vowels, is well known. If we compare the resultant clusters with the knowledge-based division of [5]

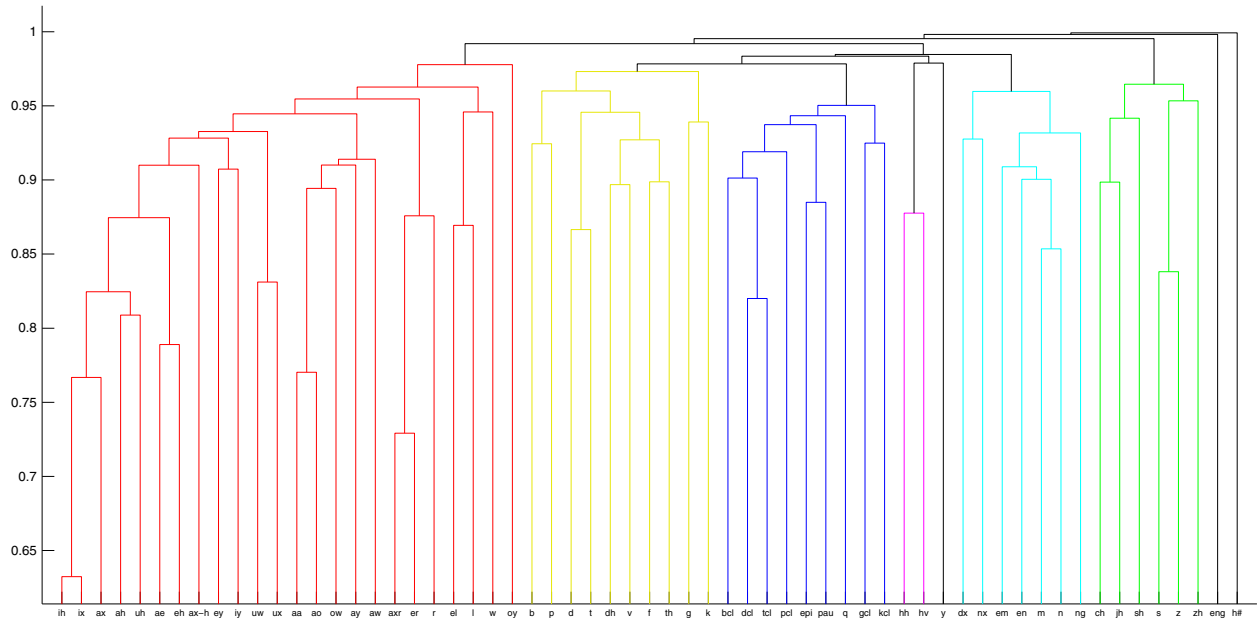


Figure 5 Hierarchical phone clustering dendrogram using the average distance between clusters (*cophenetic* correlation coefficient = 0.873).

presented in Table 2, only [y], [w] and [oy] are in a separate cluster. Nevertheless, these phones stay in a single cluster not because they are acoustically different but due to the little confusions with other phones. Another strong cluster is that of nasals. In this case, the data-driven division is the same as the knowledge-based division. Affricatives are set in a separate cluster (the knowledge-based proposal sometimes placed affricatives with stops and at others with fricatives). For fricatives and stops, the method relies on more than two clusters. The number of confusions between some fricatives and stops suggests that acoustically they exhibit similarities.

The *cophenetic* correlation coefficient referred to in Figure 5 legend is a measure of how faithfully the dendrogram preserves the pairwise distances between the original unmodelled distances [29]. The closer the value of the *cophenetic* correlation coefficient is to 1, the more accurately the clustering solution reflects the data.

Since the goal is phone recognition, the question remains as to how can these clusters help to improve phone recognition. The next section describes the followed approach—a hierarchical classification of different levels of phonetic information. Several intermediate classifiers offer posterior probability predictions for BPC, achieving phone detail at the end.

Table 1 Sixty-one TIMIT confusion-division results in terms of nine clusters

Clusters	TIMIT-labelled phones
Cluster 1	bcl dcl epi gcl kcl pau pcl q tcl
Cluster 2	b d dh f g k p t th v
Cluster 3	y
Cluster 4	hh hv
Cluster 5	dx em en m n ng nx
Cluster 6	aa ae ah ao aw ax ax-h axr ay eh el er ey ih ix iy l ow oy r uh uw ux w
Cluster 7	ch jh s sh z zh
Cluster 8	eng
Cluster 9	h#

Enhanced phone recognition

In this section, we use the broad classes generated automatically by the method proposed in the previous section in a phone recognition task. Given the difficulty of finding a threshold that leads to an optimum number of clusters, we decided to cut the tree in different places, which resulted in several sets of clusters. This procedure forms a hierarchical structure, from broad to fine phonetic detail. The combination of several levels of phonetic detail has

Table 2 Thirty-nine TIMIT knowledge-based division into five broad classes, from [5]

Broad classes	TIMIT-labelled phones
Vowels	l, r, w, y, er, ey, aw, ay, oy, ow, iy ,eh, ae, aa, uh, uw, ax, ix
Stops	p, t, k, b, d, g, jh, ch
Fricatives	s, z, zh, f, th, v, dh, hh
Nasals	m, n, ng
Silences	sil, dx

already been investigated in several studies, e.g. [5], where the outputs of four broad phonetic group classifiers (knowledge-based generated and trained separately) are combined in order to correct or enhance a phone classifier. Our proposal follows a similar approach, but by means of a hierarchical structure and with broad classes generated by the confusion-driven approach. Later, we show that this hierarchical classification takes advantage over the classifiers trained separately.

Hierarchical broad classes

We propose a hierarchical classification system that consists of a hybrid MLP/HMM, where the neural network architecture performs phone classification with a hierarchical set of broad class phonetic classifiers. The number of output layers in the MLP is the same as the cut places of the clustering tree, with the same order. Each cluster is characterized by the set of phones grouped by the tree cut and is called broad class. The broad class predictions from earlier classifiers are fed to the next ones in order to enhance the class discrimination in the current classifier. The last layer performs a 1-to-61 classification of the set of phones. All layers are trained concurrently so that, in training mode, targets are presented at all output layers.

The serial arrangement proposed provides several broad-class posteriors along with the phone posteriors. A better phone classifier may be achieved if all these posteriors are correctly combined. Previous study [8] shows that phone prediction may be more robust if class membership probabilities are weighted and combined. A method for finding the best set of weights based on discriminative training in a hybrid MLP/HMM system is described below.

Hierarchical MLP combination approach

The goal of a combination approach is to take advantage of the broad-class posteriors along with the phone posteriors in order to improve the global phone recognition performance. Our approach considers that each phone can be predicted by combining all the broad-class outputs associated with that phone, with weights differing for each phone. These weights are found by means of a discriminative training method. Each weight will be assigned to the logarithm of the network output which includes the phone. The global phone posteriors are found by combining the corresponding outputs of all output layers. The proposed combination rule is expressed by

$$\hat{P}(p_k|\mathbf{y}) = \frac{1}{Z}\exp\left(\sum_{l=1}^{N_L}\alpha_{c_k}^{(l)}\log\left(y_{c_k}^{(l)}\right)\right) \tag{9}$$

$\hat{P}(p_k|\mathbf{y})$ is the kth phone probability prediction, given the layer outputs, y, corresponding to the broad-class predictions in each of the N_L output layers (see Figure 6). $y_{c_k}^{(l)}$ and $\alpha_{c_k}^{(l)}$ are the network output and corresponding weight of layer l and index c_k, denoting the broad-class index (in layer l) to which the phone k belongs. Each phone is predicted by weighting all the class outputs associated with the phone k, $k \in \{1, .., 61\}$, which are different for each phone. Referring to Figure 6, there are four output layers with 9, 16, 40 and 61 clusters and according to Table 3 the phone [zh] has indexes 3, 9, 23 and 61 in the layers from 1 to 4. In this equation, Z is a normalization factor for the predictor $\hat{P}(p_k|\mathbf{y})$ for the 61 phones sum up to one.

The best set of weights is the one which gives the highest phone accuracy according to our hybrid MLP/

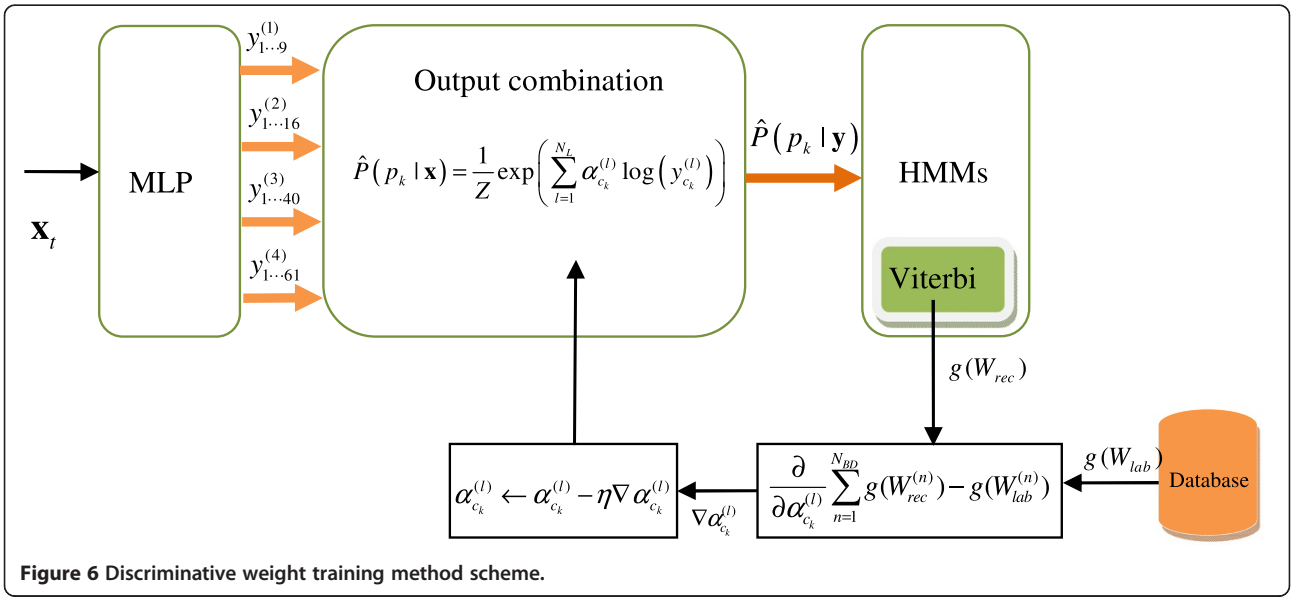


Figure 6 Discriminative weight training method scheme.

Table 3 Description of data-driven broad classes

BPC9	BPC16	BPC40
l, el, w, r, er, axr, ey, aw, ay, iy, ih, eh, ae, ah, ax, uh, ix, uw, ux, ax-h, aa, ao, ow, oy	l, el, w	l, el
		w
	r, er, axr	r, er, axr
	ey, aw, ay, iy, ih, eh, ae, ah, ax, uh, ix, uw, ux, ax-h, aa, ao, ow	ey
		aw
		ay
		iy
		ih, eh, ae, ah, ax, uh, ix
		uw, ux
		ax-h
		aa, ao, ow
	oy	oy
p, b, d, t, f, th, v, dh, k, g	p, b	p
		b
	d, t, f, th, v, dh	d, t
		f , th
		v, dh
	k, g	k
		g
jh, ch, sh, z, s, zh	jh, ch, sh	jh, ch
		sh
	z, s, zh	z, s
		zh
hh, hv	hh, hv	hh, hv
m, n, em, ng, en, nx, dx	m, n, em, ng, en	m , n
		em
		ng
		en
	nx, dx	nx
		dx
pcl, tcl, dcl, kcl, q	pcl, tcl, dcl, kcl, q	pcl
bcl, gcl, pau, epi	bcl, gcl, pau, epi	tcl, dcl
		kcl
		q
		bcl
		gcl
		pau, epi
y	y	y
eng	eng	eng
h#	h#	h#

HMM recognition system. An iterative training method based on the paradigm of discriminative training is therefore appropriate.

Discriminative training of the weights

Every kind of error should be considered (substitutions, insertions and deletions) in the definition of a cost function. Since these errors are found by the *Levenshtein* distance, the objective function should include a minimization of this distance with multiple hypotheses. However, we used a simple 1-best discriminative function, thereby avoiding the error counting, which it is better than applying the phone targets to the network output layer as is usually done. The *Levenshtein* distance aligns two label sequences. One is the reference (assumed correct) sequence, W_{lab}, and the other is the best decoding hypothesis given by the recognizer, W_{rec}. Using the Viterbi algorithm, we define an error function as

$$d\left(W_{rec}, W_{lab}\right) = g(W_{rec}) - g(W_{lab}) \tag{10}$$

where $g(W_{\text{lab}})$ and $g(W_{\text{rec}})$ represent the reference and best acoustic log likelihood of the observation sequence according to the Viterbi algorithm. This difference corresponds to a likelihood ratio. It is always greater than zero and is only zero if the two transcriptions are exactly the same (if labels and time alignments match).

If N_{BD} is the total number of training utterances, the global cost is then given by

$$E = \sum_{n=1}^{N_{\text{BD}}} d\left(W_{\text{rec}}^{(n)}, W_{\text{lab}}^{(n)}\right) \tag{11}$$

In the hybrid MLP/HMM approach, the *a priori* probability function $b_s(\mathrm{x})$ is the likelihood of observing x in the HMM state, s, being transformed in the posterior probability predicted by Equation (9).

In order to find the appropriate set of weights $\left\{\alpha_k^{(l)}\right\}$, a gradient descendent method is applied. In this case, it can be shown that the error gradient has terms of the form

$$\frac{\partial}{\partial \alpha_{c_k}^{(l)}} \log \hat{P}\left(p_k|\mathbf{y}\right) = \log\left(y_{c_k}^{(l)}\right)(1 - 1/Z) \tag{12}$$

On the other hand, the gradient of the weights is

$$\begin{aligned} \nabla \alpha_{c_k}^{(l)} &= \frac{\partial E}{\partial \alpha_{c_k}^{(l)}} \\ &= \sum_{n=1}^{N_{BD}} \left(\frac{\partial g\left(W_{rec}^{(n)}\right)}{\partial \alpha_{c_k}^{(l)}} - \frac{\partial g\left(W_{lab}^{(n)}\right)}{\partial \alpha_{c_k}^{(l)}} \right) \end{aligned} \tag{13}$$

The gradients of the log likelihoods in this expression depends on (12) because the HMM states of the path ($W_{\text{rec}}^{(n)}$ or $W_{\text{lab}}^{(n)}$) are associated with the MLP outputs, $y_{c_k}^{(l)}$.

Figure 6 illustrates the scheme of the proposed discriminative weight training method.

We used the RPROP to accelerate the convergence to a solution.

Results of the proposed hierarchical MLP architecture with broad classes found by the clustering method are presented in Section 5.3.

Experimental results

All the experiments were carried out using hybrid MLP/HMM systems. Speech was analyzed as in Section 3.1.1. Both training and testing were carried out using the TIMIT database [20] and its original 61 phoneme set. This database is often used in phone recognition benchmarking, e.g. [5,7,30]. The train and test sets correspond to the original splitting of the TIMIT database. While the training set with all si and sx sentences has 3,698 utterances, the test set consists of all si and sx sentences from the complete 168-speaker test set, which has 1,344 utterances.

The targets derive from the phone boundaries provided by the TIMIT database. Although the neural network is tailored to discriminate the full 61 TIMIT phones, these symbols are sometimes considered a too narrow description for practical use, and for evaluation purposes we collapsed the 61 TIMIT labels into the standard 39 phones as proposed by Lee and Hon [31]. In the hybrid MLP/HMM systems, the *a priori* state likelihoods are replaced by posterior probabilities, $\hat{P}(p_k|\mathbf{x})$, given directly from the MLP output layer or according to Equation (9) (Section 4.2). The performance of the MLPs was evaluated by means of a frame error rate (FER). The performance of the hybrid system was evaluated by means of Correctness (Corr) and Accuracy (Acc).

Hierarchical versus single layer classification

As described in Section 4.1, a hierarchical classification of different levels of phonetic information is proposed in order to improve phone recognition. The neural network has about 161 k parameters. It comprises eight hidden layers. The number of nodes in the layers is (in numerical order): 50-9-50-16-50-40-100-61. Even layers give posterior probability predictions for BPC and the last layer gives posterior probability predictions for phones. Thus, the proposed MLP system is trained as a function of the 61 phones and 3 additional sets of BPC, consisting of 9, 16, 40 TIMIT phone sets (viz. BPC9, BPC16, BPC40) achieved by the confusion-driven clustering approach proposed. The hierarchical phone clustering dendrogram in Figure 5 gives rise to the BPC9. The two others BPCs result from cutting the data-driven dendrogram tree into two other levels. The resulting sets of BPCs were grouped according to the division presented in Table 3.

Standard MFCC's features and derivatives are presented for input (odd) hidden layers. A context window of 290 ms was used using only 15 frame features (details in [8]).

All layers were trained concurrently so that, in training mode, targets were presented at all even layers: layers 2, 4, 6 and 8. Since even layers are trained with a *softmax* activation function, its outputs can be seen as BPC probabilities. The odd uses a sigmoid activation function. All the network weights and bias are adjusted using batch training with an RPROP algorithm [21] so as to minimize the minimum-cross-entropy error between the network output and the target values. The choice of the error function followed Bishop's suggestion [32], which was later clarified by Dunne and Campbell [33]. It states that the *softmax* activation function should couple with the cross-entropy penalty function.

In order to evaluate the performance of the proposed hierarchical network, we trained four single layer networks, each one specialized in one BPC. Each single layer network was trained with the same number of hidden nodes as the hierarchical network (e.g. BPC16 MLP has 50 hidden nodes and an output layer with 16 outputs). The results in terms of FER are given in Figure 7 and show that the hierarchical network outperformed the equivalent single layer networks with respect to BPC16, BPC40 and 61 TIMIT phones. Regarding the classification of BPC9 (9 clusters of Table 1) the performance of the single layer network is similar to the hierarchical. These results encourage the use of the hierarchical structure in further experiments. Note that these results relate to an evaluation (in terms of FER) of the test set for every training epoch.

Hierarchical confusion-driven versus hierarchical knowledge-driven phone recognition

Since confusion-driven clustering is the topic of this article, the performance of this kind of clustering has to be compared with that achieved by an expert-knowledge approach. Therefore, we trained two hierarchical MLPs from

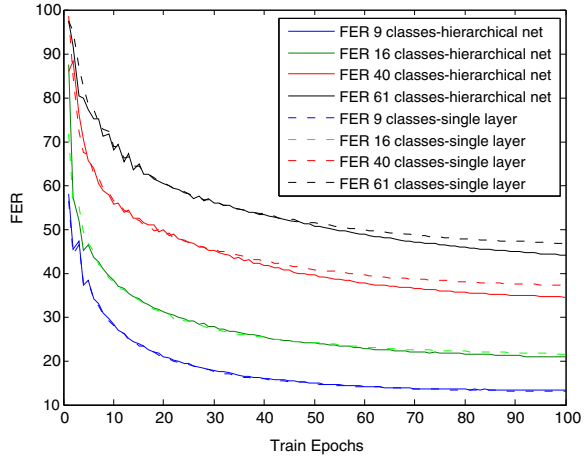


Figure 7 FER comparison of hierarchical BPC's classification and equivalent single layer BPC's classification.

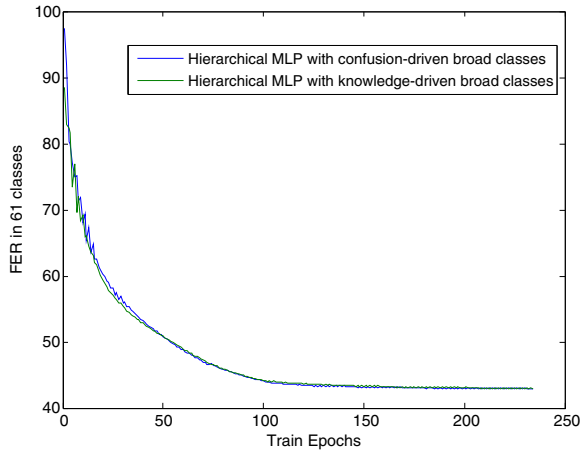


Figure 8 Comparison of FER of two similar MLPs with intermediate layers trained as a function of broad classes found by two different clustering approaches: confusion-driven and knowledge-driven.

scratch. They had a similar number of parameters, with the same number of layers but with broad classes found by means of the proposed confusion-driven clustering procedure (*Hierarchical MLP with confusion-driven broad classes*) and by the classical division based on expert knowledge (*Hierarchical MLP with knowledge-driven broad classes*). The confusion-driven broad classes division is presented in Table 3 and the knowledge-driven broad classes division is presented in Table 1 of [8]. Besides the intermediate layers having a different number of outputs, in both MLPs the output layer is trained as a function of the 61 TIMIT original phones. Figure 8 presents the comparison in terms of FER within the 39 phone set evaluation. Results for the test set are very similar in the two figures, which indicates that the proposed confusion-driven phone clustering method may be used instead of the classical methods based on human knowledge.

Phone recognition with MLP combination

One question remains unanswered: how can confusion-driven phone clustering enhance phone recognition? In view of this, we tested a hybrid MLP/HMM system with the MLP following the hierarchical structure described in Sections 3 and 4. Results are presented in Table 4. The first line shows the baseline system for which only the output layer of the MLP was considered, which gives 61 phone class membership predictions in the HMM, while the second line gives the results of the hierarchical classification structure with the outputs of each level combined by the discriminative weight training method. The results rise from a correctness rate of 71.8–73.8% which is a relative improvement of 2.8%. Regarding accuracy, the relative improvement was 2%, to reach 70%. As expected, a combination of broad-class posteriors with phone posteriors can be effective for enhancing both the correctness and accuracy rates in phone recognition.

Table 4 TIMIT phone recognition results

	%Corr	%Acc	% Improvement	
			Corr	Acc
Baseline (PP)	71.8	68.6	-	-
CDPC + PP	73.8	70.0	2.8	2.0

Baseline: considering only phone posteriors (PP) in the decoding process. CDPC + PP: the discriminative combination of the proposed confusion-driven phone clusters (CDPC), with phone posteriors (PP).

Contextualizing with TIMIT state-of-art results

The results presented here are not comparable with those published in [34] because the authors of that study evaluated their system with phone classification and not phone recognition, as we have done. But the results compare favourably with the findings presented by an ASAT (Automatic Speech Attribute Transcription) group in [35] and by Morris and Fosler-Lussier [7]. The only factor those studies have in common with this study is that they present results for the same conditions (same speech material and same recognition rates). The ASAT group [35] uses confidence scores of phonetic attributes classes given by an MLP, an HMM and an SVM in a CRF for phone recognition. They indicate an accuracy rate of 69.52%. This value is lower than our CDPC + PP results. Morris and Fosler-Lussier [7] use phonological features provided by an ANN together with 61 class posteriors, provided by another ANN, also as input of a CRF, and achieve a 71.49% accuracy rate. Our results are close, which shows that the very different approaches provide similar performance results. It would appear that higher results will not be achieved for the TIMIT corpus unless other approaches are used, which should include a widening of the phonetic context as was successfully done in [6]. Milestones in phone recognition using the TIMIT database can be found in [30].

Conclusions

This article has described a phone clustering method based on a phone similarity metric which is computed from a confusion matrix generated from the output of a phone recognition system. The results show that equally good performances are achieved using the broad classes given by the proposed clustering method and using the broad classes defined using human knowledge. In addition to the proposed method overcoming the subjectivity and time-consumption related to the human-based on, employing a clustering method like this make possible the creation of broad classes also in systems where the recognition unit is not the phone. It is also much easier to modify the speech database phone (or

other unit) set. This may be necessary in cases where the speech databases are not extensive enough to have a sufficient number of occurrences of all phones. Another advantage of the method is that the approach proposed in this article can easily be generalized and applied to other languages or to multilingual systems. It should be noted that the proposed method is highly dependent on the performance of the starting recognizer. The quality of the acoustic models of the phone recognizer must be good enough to provide a warrantable confusion matrix.

A hierarchical structure for training the intermediate broad classes has also been proposed. It has been found to be superior to training each broad class set separately, as is usually done. Furthermore, the scope of the current proposal extended to applying the confusion-driven phone clustering method to a phone recognition system. In view of this, a hierarchical structure was built where middle levels were trained as a function of the clusters generated by the proposed clustering procedure, to arrive at phone detail in the last level. Enhanced phone recognition was in fact found by combining the contribution of each cluster of each level in the final phone posterior estimation. Since the optimal combination depends on each individual phone, an original discriminative method was used to find the best set of weights for each phone. The results were encouraging as the correctness and accuracy rates exhibited relative improvements of 2.8 and 2%, respectively, relatively to a baseline system.

Better results might be obtained if a new confusion matrix were computed from the output of this improved phone recognition system; the confusion-driven phone clustering method can then be applied once more, with the result that better cluster divisions are naturally found, and the phone discrimination is thus improved yet again.

Endnote

[a]"Confusion-driven" is used since the phone clustering is a function of a confusion matrix.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Carla Lopes would like to thank the Portuguese foundation: Fundação para a Ciência e a Tecnologia for the PhD Grant (SFRH/BD/27966/2006).

Author details

[1]Instituto de Telecomunicações, Polo II, Coimbra P-3030-290, Portugal. [2]ESTG, Instituto Politécnico de Leiria, Campus 2, Leiria P-2411-901, Portugal. [3]Universidade de Coimbra – DEEC, Pólo II, Coimbra P-3030-290, Portugal.

Received: 1 April 2011 Accepted: 17 May 2012
Published: 23 July 2012

References

1. T. Kempton, R.K. Moore, *Language identification: insights from the classification of hand annotated phone transcripts, in Proc* (Odyssey Workshop on Speaker and Language Recognition, Stellenbosch, South Africa, 2008).
2. J. Yuan, M. Liberman, *Robust speaking rate estimation using broad phonetic class recognition* (Proceedings of ICASSP2010, Dallas, USA, 2010), pp. 4222–4225.
3. L. Yang, J. Zhang, Y. Yan, *Acoustic units selection in Chinese–English bilingual speech recognition, in Proc Non Linear Speech Processing* (Paris, France, 2007).
4. A. Zgank, B. Horvat, Z. Kacic, Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. Speech Commun. **47**(3), 379–393 (2005).
5. P. Scanlon, D. Ellis, R. Reilly, Using broad phonetic group experts for improved speech recognition. IEEE Trans. Audio Speech Lang. Process. **15**(3), 803–812 (2007).
6. S.M. Siniscalchi, P. Schwarz, C.-H. Lee, *High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring*, vol. 4 (Proc. ICASSP, Honolulu, USA, 2007), pp. 869–872.
7. J. Morris, E. Fosler-Lussier, Conditional random fields for integrating local discriminative classifiers. IEEE Trans. Acoust. Speech Lang. Process. **16**(3), 617–628 (2008).
8. C. Lopes, F. Perdigão, *A hierarchical broad-class classification to enhance phoneme recognition* (Proc. EUSIPCO-2009, Glasgow, UK, 2009), pp. 1760–1764
9. T.N. Sainath, V. Zue, *A comparison of broad phonetic and acoustic units for noise robust segment-based speech recognition* (Proc. Interspeech, Brisbane, Australia, 2008), pp. 2378–2381.
10. A. Halberstadt, J. Glass, *Heterogeneous acoustic measurements for phonetic classification* (Proc. Eurospeech, Rhodes, Greece, 1997), pp. 401–404.
11. R.C. Rose, P. Momayyez, Integration of multiple feature sets for reducing ambiguity in ASR, in *Proc International Conference on Acoustics, Speech, and Signal Processing-ICASSP2007*, Honolulu. USA **4**, 325–328 (April 2007).
12. A. Juneja, C. Espy-Wilson, *Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning*, vol. 2 (Proc. International Conference on Neural Information Processing ICONIP, Singapore, 2002), pp. 726–730.
13. A.M.A. Ali, J. Van, P. der Spiegal, P. Mueller, Acoustic-phonetic features for the automatic classification of stop consonants. IEEE Trans. Speech Audio Process. **9**(8), 833–841 (2001).
14. A.M.A. Ali, J. Van der Spiegel, P. Mueller, *An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants* (Proc. International Conference on Acoustics, Speech, and Signal Processing - ICASSP98, Seattle, USA, 1998), pp. 961–964.
15. M. Brian, B. Etienne, *Phone clustering using the Bhattacharyya distance* (Proc. International Conference on Spoken Language Processing, Philadelphia, USA, 1996), pp. 2005–2008.
16. B. Imperl, Z. Kacic, B. Horvat, A. Zgank, Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones. Speech Commun. **39**(3–4), 353–366 (2003).
17. J. Köhler, *Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds* (Proc. International Conference on Spoken Language Processing, Philadelphia, PA, USA, 1996), pp. 2195–2198.
18. C. Chelba, R. Morton, *Mutual information phone clustering for decision tree induction* (Proc. International Conference on Spoken Language Processing, Denver, USA, 2002), pp. 1005–1008.
19. P. Mareuil, C. Corredor Ardoy, M. Adda-Decker, *Multi-lingual automatic phoneme clustering* (Proc. 14th International Conference on Phonetic Science, ICPhS-99, San Francisco, 1999), pp. 1209–1212.
20. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, *DARPA* (TIMIT acoustic-phonetic continuous speech corpus CD-ROM, NIST, 1990).
21. M. Riedmiller, H. Braun, *A direct adaptive method for faster backpropagation learning: the RPROP algorithm* (Proc. International Conference on Neural Networks, Francisco, CA, 1993), pp. 586–591.
22. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, (The HTK Book. Revised for HTK Version 3.4, Cambridge University Engineering Department, Cambridge, 2006).
23. C. Lopes, F. Perdigão, *Improved performance evaluation of speech event detectors* (Proc. Interspeech2006, Pittsburgh, USA, 2006), pp. 2190–2193.
24. D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge University Press, UK, 1997).
25. M. Vihola, M. Harju, P. Salmela, J. Suontausta, J. Savela, *Two dissimilarity measures for HMMs and their application in phoneme model clustering*, vol. 1 (Proc. ICASSP 2002, Orlando, FL, 2002), pp. 933–936.
26. M. Vihola, *Dissimilarity measures for hidden Markov models and their application in multilingual speech recognition* (MSc thesis, Tampere University of Technology, May 2002).

27. W. Day, H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods. J. Classif **1**, 7–24 (1984).
28. S. Johnson, Hierarchical clustering schemes. Psychometrika **2**, 241–254 (1967)
29. R.R. Sokal, F.J. Rohlf, The comparison of dendrograms by objective methods. Taxon **11**, 33–40 (1962).
30. C. Lopes, F. Perdigão, *Phone Recognition on the TIMIT Database - Chapter in Speech Technologies* ((In-Tech, 2011). ISBN 978-953-307-996-7.
31. K. Lee, H. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **37**(11), 1642–1648 (1989).
32. C. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).
33. R. Dunne, N. Campbell, *On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function* (Proc. 8th Australian Conf. on Neural Networks, Melbourne, Australia, 1997), pp. 181–185.
34. A. Gunawardana, M. Mahajan, A. Acero, J. Platt, *Hidden conditional random fields for phone classification* (Proc. Interspeech2005, Lisbon, Portugal, 2005), pp. 1117–1120.
35. I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. Siniscalchi, Y. Tsao, Y. Wang, *Detection-based ASR in the automatic speech attribute transcription project* (in Proc. of Interspeech2007, Antwerp, Belgium, 2007), pp. 1829–1832.

doi:10.1186/1687-6180-2012-158
Cite this article as: Lopes and Perdigão: **Broad phonetic class definition driven by phone confusions.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:158.