# A Clustering Algorithm for the Fast Match of Acoustic Conditions in Continuous Speech Recognition

L.J. Rodríguez and M.I. Torres*

Pattern Recognition & Speech Technology Group
DEE. Facultad de Ciencia y Tecnología. Universidad del País Vasco
Apartado 644. 48080 Bilbao. SPAIN
e-mail: luisja@we.lc.ehu.es

**Abstract.** In practical speech recognition applications, channel/environment conditions may not match those of the corpus used to estimate the acoustic models. A straightforward methodology is proposed in this paper by which the speech recognizer can match the acoustic conditions of input utterances, thus allowing instantaneous adaptation schemes. First a number of clusters is determined in the training material in a fully unsupervised way, using a dissimilarity measure based on shallow acoustic models. Then accurate acoustic models are estimated for each cluster, and finally a fast match strategy, based on the shallow models, is used to choose the most likely acoustic condition for each input utterance. The performance of the clustering algorithm was tested on two speech databases in Spanish: SENGLAR (read speech) and CORLEC-EHU-1 (spontaneous human-human dialogues). In both cases, speech utterances were consistently grouped by gender, by recording conditions or by background/channel noise. Furthermore, the fast match methodology led to noticeable improvements in preliminary phonetic recognition experiments, at 20-50% of the computational cost of the ML match.

## 1 Introduction

One of the most challenging issues posed by current applications of continuous speech recognition is the increased acoustic variability due to spontaneous speech, speaker features, channel or environmental conditions, etc. Many adaptation techniques have been proposed to increase the robustness of speech recognizers to speaker features and mismatched environment conditions [1]. One of them consists of organizing the training material into clusters of acoustically similar utterances, then training specific acoustic models for them, and finally matching the acoustic conditions (i.e. the most suitable cluster) for each input utterance.

* This work was partially supported by the University of the Basque Country, under grant 9/UPV 00224.310-13566/2001, and the Spanish MCYT, under project TIC2002-04103-C03-02.

The training material may be clustered in a supervised way by using *a priori* knowledge about speaker identities or environmental conditions of utterances [2]. But in practical applications such knowledge might be unreliable or unavailable. In this framework, an unsupervised clustering algorithm is needed to automatically determine an optimal partition in the set of utterances, as some authors have proposed [3–5].

In a previous study, we developed a clustering algorithm to find an optimal partition of speakers in the training material. Then we trained speaker-class models and during recognition the most suitable speaker classes for each input utterance were selected or combined in a fast and straightforward manner, using shallow acoustic models [6]. In that study, all the samples from any given speaker had to be moved to the same speaker-class. Here we apply the same methodology but in a fully unsupervised way: information about speaker identity is left out and each utterance is moved independently.

Assuming a non homogeneous set of speech utterances in the training corpus, we propose an unsupervised clustering algorithm which automatically finds an *optimal* partition in that set, using a dissimilarity measure based on shallow acoustic models. Once the optimal partition is defined, hidden Markov models (HMM) are estimated for each cluster. During recognition, the shallow models are applied to the input utterances in a straightforward manner, without recognizing them, to choose the most suitable clusters. The corresponding HMMs are then applied to get a number of decodings for each input utterance, and finally the most likely string is hypothesized. The number of decodings actually done depends on the *sharpness* of the decision, i.e. on the number of cluster candidates. In the best case, a single decoding would be carried out for each input utterance. Assuming that each cluster represents specific acoustic conditions (a pool of gender, channel and environment), this procedure can be viewed as a fast match of acoustic conditions. The fast match strategy is critical to making cluster models useful in actual applications, since the *Maximum Likelihood* (ML) match —i.e. carrying out all the decodings, one for each cluster, and then selecting the decoded string with the highest likelihood— would be too costly.

The rest of the paper is organized as follows: Section 2 describes the histogram models used to represent the clusters; Section 3 briefly outlines the clustering algorithm; Section 4 describes the fast match approach followed in this study; Section 5 presents experimental evaluation of the clustering algorithm on two speech databases in Spanish, and phonetic recognition results which provide evidence of the usefulness of the fast match strategy; finally, Section 6 summarizes the main contributions of the study.

## 2 The Shallow Acoustic Model

Let $M$ be the number of acoustic vectors used to represent the speech signal at each time $t$. Then each sample $X(t)$ consists of $M$ vectors, $X_j(t)$ with $j = 1, \ldots, M$. First, Vector Quantization (VQ) is applied to build a *codebook* of $N$ centroids for each acoustic representation. These codebooks minimize the average distortion in quantifying the acoustic vectors of the training corpus. Once the

VQ codebooks are defined, each vector $X_j(t)$ can be replaced by a single symbol $Y_j(t) \in \{1, \ldots, N\}$, corresponding to the index of the nearest centroid.

Now, assuming that the training corpus is partitioned into $S$ clusters, consider the cluster $i$, for which $c(i)$ samples are available. We store in $c(k, j, i)$ the number of times $Y_j(t) = k$ in the set of samples corresponding to the cluster $i$, and define the discrete distribution $P_j(k|i)$ as:

$$P_j(k|i) = \frac{c(k, j, i)}{c(i)} \ . \tag{1}$$

This is an empirical distribution based on the histograms of the symbols at each acoustic stream. Hereafter, we will refer to it as *histogram model.* Note that for any $j$ $\sum_{k=1}^{N} c(k, j, i) = c(i)$, so that $\sum_{k=1}^{N} P_j(k|i) = 1$. The probability that a quantified speech sample $Y(t)$ is produced in the acoustic conditions represented by cluster $i$ is defined as the joint discrete distribution:

$$P(Y(t)|i) = \prod_{j=1}^{M} P_j(Y_j(t)|i) \ . \tag{2}$$

Finally, the probability that a speech utterance $Y = \{Y(t)|t = 1, \ldots, T\}$ is produced in the acoustic conditions represented by cluster $i$ is computed as follows:

$$P(Y|i) = \prod_{t=1}^{T} P(Y(t)|i) \ . \tag{3}$$

## 3 The Clustering Algorithm

A top-down clustering scheme was applied starting from a single cluster, iteratively splitting one of the clusters and readjusting the allocation of utterances until not enough speech frames were available or the average distortion decreased below a certain threshold.

Before writing the algorithm, we must give some definitions. First, a histogram model is constructed for each speech utterance $l$, based on the set of quantified samples corresponding to that utterance, with $\Upsilon(l) = \{Y(t)|t = 1, \ldots, s(l)\}$, $s(l)$ being the length of the utterance. Then the *dissimilarity* of $l$ with regard to a given cluster $i$, $d(l; i)$, is defined as follows:

$$d(l; i) = -\log \left\{ \frac{P(\Upsilon(l)|i)}{P(\Upsilon(l)|l)} \right\} \ , \tag{4}$$

where $P(\Upsilon(l)| \cdot)$ is computed as the joint probability of all the quantified speech samples corresponding to the utterance $l$, given a histogram model (equation 3).

At any iteration $n$ of the clustering algorithm, each utterance $l$ is assigned to the *closest* cluster $i_n^{(l)}$ in the partition $\Pi(n)$: $i_n^{(l)} = \arg\min_{g \in \Pi(n)} d(l; g)$. Taking this into account, the distortion of $\Pi(n)$ is defined as:

$$R(n) = \frac{1}{L} \sum_{l=1}^{L} d(l; i_n^{(l)}) = -log \left[ \prod_{l=1}^{L} \frac{P(\Upsilon(l)|i_n^{(l)})}{P(\Upsilon(l)|l)} \right]^{1/L} \tag{5}$$

where $L$ is the number of speech utterances in the training corpus.

Finally, for each cluster $i$, the first and second centroid utterances, $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$, are defined as those yielding the two smallest values of the dissimilarity with regard to that cluster:

$$\gamma_1^{(i)} = \arg\min_{l \in i} d(l;i) \;\; ; \quad \gamma_2^{(i)} = \arg\min_{l \in i, l \neq \gamma_1^{(i)}} d(l;i) \tag{6}$$

The clustering algorithm is described in detail in the following paragraphs:

```
1. For each utterance $l \in \{1,\ldots,L\}$ and for each acoustic stream $j \in \{1,\ldots,M\}$, the utterance histograms $s(k,j,l)$ are counted, and the normalizing factor $s(l) = \sum_{k=1}^{N} s(k,1,l)$ computed.
2. Initially ($n = 0$), a single cluster is defined ($S = 1$) including all the utterances: $\forall l$, $i_0^{(l)} = 1$. The clustering distortion $R(0)$ is computed. Also, for each acoustic representation $j \in \{1,\ldots,M\}$ the histogram model of the initial cluster is computed as follows: $c(k,j,1) = \sum_{l=1}^{L} s(k,j,l)$ and $c(1) = \sum_{l=1}^{L} s(l)$.
3. repeat
   3.1 $n \leftarrow n + 1$
   3.2 For each cluster $g \in \Pi(n)$, obtain the first and second centroid utterances, $\gamma_1^{(g)}$ and $\gamma_2^{(g)}$, and the average cluster distortion, computed as $D(g) = \frac{1}{L(g)} \sum_{l \in g} d(l;g)$, where $L(g)$ is the number of speech utterances in $g$. Add this information to a list of cluster split candidates, $c_{cand}$, in descending order of $D(g)$.
   3.3 while $c_{cand} \neq \emptyset$ do
      3.3.1 Extract the first item of the list: $(g, \gamma_1^{(g)}, \gamma_2^{(g)})$, and split cluster $g$ in two, taking as seed models of the new clusters those of $\gamma_1^{(g)}$ and $\gamma_2^{(g)}$, respectively.
      3.3.2 repeat
               - For each utterance $l$, assign it to the nearest cluster
               - For each cluster $i$, recompute the histogram model using the counts $s(k,j,l)$ and $s(l)$ of the utterances assigned to it.
            until maximum number of iterations or clusters unchanged
      3.3.3 if the new partition is valid then
               { $S \leftarrow S + 1$; Compute $R(n)$; Empty $c_{cand}$; }
            else
               { Recover the partition at $n - 1$; $R(n) \leftarrow R(n-1)$; }
   until $(R(n-1) - R(n))/R(n) < \tau$
4. Store the partition information and the corresponding histogram models.
```

In the above algorithm $\tau > 0$ is an empirical threshold for the relative decrease in average distortion. Also, each time a new partition is generated, all the clusters must contain a minimum number of speech frames to guarantee the trainability of the acoustic models. When not enough frames are available for any of the clusters, the previous partition is recovered and another splitting explored (step 3.3.3). Note also that the candidate splittings are explored in descending order of $D(g)$, so that the cluster with the highest distortion is split first.

## 4 The Fast Match Strategy

During recognition, the most suitable acoustic model(s) must be selected/combined for each input utterance. Various alternatives were explored in a previous study, where each cluster represented a speaker class [6]. The *Maximum likelihood* (ML) match approach, consisting of carrying out $S$ decodings, one for each HMM set, and selecting the one that yielded the highest likelihood, was found to be the optimal but also the most expensive alternative. On the other hand, if the histogram models were used to *pre-select a beam of candidates* —thus drastically reducing the number of decodings—, the same performance was obtained at a much lower cost. In practice, the average number of decodings was reduced to around two or three.

Taking these results into account, for each input utterance we have considered only those clusters whose histogram probabilities are higher than a heuristically fixed threshold (70% of the maximum value). Decodings are obtained only for them, and finally the decoded string that yields the highest likelihood is hypothesized. This is a kind of beam selection, motivated by the fact that sometimes the most suitable cluster —in terms of acoustic likelihood— yields histogram probabilities near but below the maximum.

## 5 Experimental Results

### 5.1 Databases

A phonetically and gender-balanced read speech database in Spanish, called SENGLAR, acquired at 16 kHz in laboratory conditions, was considered in the first place to tune the clustering algorithm. The training corpus consisted of 1529 utterances, pronounced by 57 (29 male, 28 female) speakers, and included 60399 phone samples with a total duration of around 80 minutes. The test corpus consisted of 700 utterances, pronounced by 33 (18 male, 15 female) speakers, and included 32034 phones with a total duration of around 40 minutes.

A spontaneous speech database in Spanish called CORLEC-EHU-1 [7], composed of 42 human-human dialogues taken from radio and TV broadcasts using an analog tape recorder, was considered in the second place to test the proposed methodology in more difficult conditions: variable and noticeable background/channel noise, presence of spontaneous speech events, pronunciation variability, etc. The training corpus consisted of 1421 utterances, pronounced by 67 (49 male, 18 female) speakers, and included 187675 phone samples with a total duration of around 225 minutes. The test corpus consisted of 704 utterances, pronounced by 35 (21 male, 14 female) speakers, and included 93415 phones with a total duration of around 114 minutes.

### 5.2 Results of Clustering

The mel-scale cepstral coefficients (MFCC) and energy (E) —computed in frames of 25 milliseconds, taken each 10 milliseconds— were used as acoustic features. The first and second derivatives of the MFCCs and the first derivatives of E were also computed. Four acoustic streams were defined: MFCC, $\Delta$MFCC, $\Delta^2$MFCC and (E,$\Delta$E). Finally, the LBG vector quantization algorithm [8] was applied to get four codebooks, each one consisting of 256 centroids.

The clustering algorithm was run using the training corpora of the two databases described above. At least 30000 speech frames (5 minutes) were required for each cluster to be valid. The maximum number of convergence iterations (step 3.3.2) was set at 20, and the threshold for the relative decrease in average distortion was set at $\tau = 0.01$. This resulted in 8 clusters for SENGLAR and 17 clusters for CORLEC-EHU-1.

SENGLAR was built by integrating three sub-corpora, called *FRASES*, *EUROM1* and *PROBA*, recorded in different places with slightly different hardware, so that not only speaker characteristics but also channel features may differ from one utterance to other. As shown in Table 1, all the clusters except for #3 and #4 consisted of utterances from one single sub-corpus. Additionally, clusters were formed almost exclusively either by male or by female speakers. This means that channel and speaker characteristics were effectively working to separate clusters from one another.

With regard to CORLEC-EHU-1, besides gender, two channel/environment conditions were clearly separated by the clustering algorithm: radio and TV interviews. In fact, 13 of the 17 clusters were pure in terms of gender and channel/environment, which represents 51.76% of the training frames. The remaining 4 clusters consisted of a pool of male/female, radio/TV utterances.

**Table 1.** Distribution of speech utterances after clustering in SENGLAR.

| | **FRASES** | | **EUROM1** | | **PROBA** | |
|---|---|---|---|---|---|---|
| | **male** | **female** | **male** | **female** | **male** | **female** |
| **Cluster #1** | 0 | 0 | 120 | 8 | 0 | 0 |
| **Cluster #2** | 119 | 0 | 0 | 0 | 0 | 0 |
| **Cluster #3** | 0 | 0 | 302 | 2 | 60 | 0 |
| **Cluster #4** | 0 | 0 | 1 | 6 | 14 | 100 |
| **Cluster #5** | 0 | 0 | 24 | 236 | 0 | 0 |
| **Cluster #6** | 0 | 262 | 0 | 0 | 0 | 0 |
| **Cluster #7** | 0 | 0 | 0 | 143 | 0 | 0 |
| **Cluster #8** | 132 | 0 | 0 | 0 | 0 | 0 |

### 5.3 Phonetic Recognition Results

Phonetic recognition experiments were carried out using the HMMs obtained through the unsupervised clustering methodology described above. MAP estimates were applied to get more robust models (only the Gaussian means and weights were re-estimated) [9]. During recognition, the fast match strategy described in Section 4 was applied. In the case of SENGLAR, the set of context-independent sublexical units consisted of 23 phone-like units (PLUs) plus one extra unit for *silences*. In the case of CORLEC-EHU-1, besides the 23 PLUs 14 extra units were defined to model spontaneous speech events such as noises, lengthenings, filled pauses, silent pauses, etc. A set of left-side biphones was also defined in both cases, taking into account only the trainability of the corresponding models (at least 300 training samples were required). Left-side biphones were applied jointly with context-independent units to guarantee acoustic coverage. Each sublexical unit was represented with a left-right Continuous-Density HMM consisting of three states with self-loops but no skips. Phonological restrictions were applied only when dealing with left-side biphones. Finally, the extra units representing spontaneous speech events were either filtered or mapped into PLUs before the recognized and the correct strings were aligned. Phonetic recognition rates obtained using HMMs adapted through unsupervised clustering are shown in Table 2. To allow suitable comparisons, results using non-adapted HMMs (estimated using the whole training corpus) and HMMs adapted through speaker clustering [6] are also shown.

**Table 2.** Phonetic recognition rates obtained using non-adapted HMMs and HMMs adapted through speaker clustering and unsupervised clustering of utterances, for SENGLAR and CORLEC-EHU-1. Experiments were carried out using context-independent (CI) and context-dependent (CD) sublexical units.

| | **SENGLAR** | | **CORLEC-EHU-1** | |
|---|---|---|---|---|
| | **CI** | **CD** | **CI** | **CD** |
| **Non-adapted HMMs** | 72.38 | 75.38 | 52.42 | 57.09 |
| **Adapted HMMs: Speaker Clustering** | 74.41 | 75.79 | 53.89 | 58.05 |
| **Adapted HMMs: Unsupervised Clustering** | **74.33** | **75.78** | **53.53** | **57.58** |

The HMMs adapted through unsupervised clustering outperformed the non-adapted HMMs in all cases. In the case of SENGLAR, improvements were quite noticeable when using context-independent models (7.06% relative error reduction), whereas only slight imoprovements were achieved with context-dependent models (1.62% relative error reduction). This is probably due to a lack of samples for the context-dependent models. In the case of CORLEC-EHU-1 more training samples were available, but the higher acoustic variability of spontaneous speech and especially the adverse channel/environment conditions made the improvements smaller in both cases (2.33% and 1.17% relative error reduction, respectively). In fact, phonetic recognition rates for CORLEC-EHU-1 are around 20 absolute points lower than those obtained for SENGLAR. So, though the usupervised clustering of utterances helps in modeling channel/environment variabilities, more specific strategies (noise compensation techniques, noise robust features, etc.) seem to be needed. On the other hand, the performance attained through unsupervised clustering is almost the same as that obtained through speaker clustering, with no information about either speaker identities or channel/environment conditions. Finally, the average number of decodings in the fast match was 4.09 in the case of SENGLAR and 3.64 in the case of CORLEC-EHU-1, which works out at 51.13% and 21.41% of the computational cost of the ML match, respectively.

## 6 Concluding remarks

A new clustering algorithm is presented in this paper which automatically determines an optimal partition in the training corpus of a speech database using a dissimilarity measure based on shallow acoustic models. Then accurate acoustic models are estimated for each cluster, which represent specific (but unknown) speaker/environment conditions. During recognition, the most suitable clusters are selected using a fast match strategy, combining acoustic probabilities computed with the shallow models and full decodings obtained with HMMs. Preliminary results are presented for two databases of read and spontaneous speech in Spanish, revealing that speaker and channel/environment characteristics are implicitly taken into account by the clustering algorithm. A 7% decrease in error rate was attained in phonetic recognition experiments over read speech, at half the computational cost of the ML match. For spontaneous speech, the relative error decrease was slightly higher than 2%, at 20% of the cost of the ML match. Our current work involves applying this methodology to larger corpora of non homogeneous speech, such as those recorded in human-machine dialogue tasks. Note that unsupervised adaptation to speaker and environment conditions is crucial to increasing the robustness of spoken dialogue systems.

## References

1. Gales, M.J.F.: Adaptive Training for Robust ASR. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Madonna di Campiglio (Italy) (2001)
2. Gao, Y., Padmanabhan, M., Picheny, M.: Speaker Adaptation Based on Pre-Clustering Training Speakers. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). (1997) 2091–2094
3. Jin, H., Kubala, F., Schwartz, R.: Automatic Speaker Clustering. In: Proceedings of the DARPA Speech Recognition Workshop, Chantilly, VA (1997) 108–111
4. Chen, S.S., Gopalakrishnan, P.S.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA (1998)
5. Ajmera, J., Wooters, C.: A Robust Speaker Clustering Algorithm. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), St. Thomas, U.S. Virgin Islands (2003)
6. Rodríguez, L.J., Torres, M.I.: A Speaker Clustering Algorithm for Fast Speaker Adaptation in Continuous Speech Recognition. In Sojka, P., Kopeček, I., Pala, K., eds.: Proceedings of the 7th International Conference on Text, Speech and Dialogue (TSD 2004). Lecture Notes in Artificial Intelligence LNCS/LNAI 3206, Brno, Czech Republic, Springer-Verlag (2004) 433–440
7. Rodríguez, L.J., Torres, M.I.: Annotation and Analysis of Acoustic and Lexical Events in a Generic Corpus of Spontaneous Speech in Spanish. In: Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo Institute of Technology, Tokyo, Japan (2003) 187–190
8. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications **28** (1980) 84–95
9. Gauvain, J.L., Lee, C.H.: Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing **2** (1994) 291–298