Inter-Phone and Inter-Word Distances for Confusability Prediction in Speech Recognition

Jan Anguita and Javier Hernando
TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034, Barcelona, Spain
{jan,javier}@talp.upc.es

Resumen: En este trabajo se investigan nuevas distancias entre fonemas y entre palabras que se han usado para predecir si dos palabras del vocabulario de un sistema de reconocimiento del habla se van a confundir o no. La distancia entre palabras se calcula a partir de un alineamiento entre las transcripciones fonéticas de las palabras sumando las distancias entre los fonemas alineados. Se propone una nueva solución donde la distancia entre fonemas usada para alinear no es la misma que la que se usa para calcular la distancia entre palabras. La primera está basada en conocimiento fonético. La segunda se obtiene a partir de los modelos acústicos de los fonemas con una nueva fórmula que proponemos. También se han usado dos tipos de alineamientos: con o sin inserciones y omisiones. Para evaluar la predicción se han calculado las tasas de falso rechazo y falsa aceptación y se ha obtenido un Equal Error Rate de menos del 2%.
Palabras clave: Distancia entre fonemas, distancia entre palabras, predicción, confusión.

Abstract: In this work we investigate new inter-phone and inter-word distances and we apply them to predict if two words of the lexicon of an Automatic Speech Recognition (ASR) system are likely to be confused. The inter-word distance is calculated from an alignment between the phonetic transcriptions of the words by adding the distances between the aligned phones. We bring a new solution in which the inter-phone distance used for computing the inter-word distance is not the same used to compute the phonetic alignment. The first one is calculated between the acoustic models of the phones with a new formula that we propose. The second one is based on phonetic knowledge. We also use two different kinds of alignments: either with or without insertions and deletions. In order to evaluate the performances, we introduce a classical false acceptance/false rejection framework and the prediction Equal Error Rate (EER) was measured to be less than 2%.
Keywords: Inter-phone distance, Inter-word distance, confusability prediction.

1 Introduction

Distance measures between phones or between words are important in some applications of phonology, such as the alignment of phonetic sequences (Covington, 1996; Somers, 1999; Kondrak, 2000); or applications of Automatic Speech Recognition (ASR), such as the selection of the lexicon. In the literature we can find several proposals of distances between words in order to help to design the lexicon of an ASR system so that its words are as less confusable as possible (Tan et al., 1999; Roe and Riley, 1994; Pouységur, 2001). In this work we go a step further and we propose to classify the word pairs into two classes: confusable or not confusable, i.e., if they are likely to be confused by an ASR system or not. This approach provides a powerful tool since, if two words of the lexicon of an ASR system are confusable, it will warn the person who is designing it, giving him the possibility of changing one of them for a synonym. In this way, the application remains the same but the probability of confusion decreases.

In order to do this classification, we calculate a distance between the phonetic transcriptions of the word pair and we classify it as confusable or not confusable using a threshold. This distance is based on a new algorithm where the phonetic transcriptions of the two words are aligned using Dynamic Programming (Wagner and Fischer, 1974) and, after, the inter-word distance is calculated as the sum of the distances between the aligned phones. The new proposal is that the inter-phone distances used to do the alignment and the ones used to calculate the inter-word distance once the alignment is done, are not the same. The first one is based on phonetic knowledge, whereas the second one is obtained by calculating a new distance between the acoustic models (Hidden Markov Models) of the phones. We have evaluated the performances of these distances with two different kinds of alignments: either with or without insertions and deletions.

In order to evaluate the performance, we introduce a classical false acceptance/false rejection framework for comparing a posteriori classification obtained by testing ASR systems with the a priori classification produced by the method. To obtain data to test is not a trivial problem since the confusability of two words depends on the whole system: the rest of the lexicon, the kind of used models, etc., but we make a proposal to solve it.

The organization of this paper is as follows. In section 2 the new inter-phone distances are presented. In section 3 the two different kinds of alignments used in this work are described. In section 4 the classical DTW distance is reviewed and, the new distance, called Phonetic Acoustic Dissimilarity measure (PAD), and the classification procedure are introduced. In section 5 the experiments and the results are presented and, finally, section 6 concludes the paper.

2 Inter-Phone distances

In this section we present two different kinds of inter-phone distances: one is calculated from the acoustic models of the phones, and the other one is based on phonetic knowledge.

2.1 Inter-Phone Distance Between Acoustic Models

One way to obtain a measure of distance between two phones is to calculate the distance between its acoustic models (Tan et al., 1999). Since in modern ASR systems the acoustic units are usually modelled by Hidden Markov Models (HMM) (Rabiner, 1989), in this paper we propose the following distance measure between the HMMs of two phones:

$$d_{HMM}(p_1,p_2)=\begin{cases}\dfrac{\sum_{Q}\left(P(Q)\frac{1}{L}\sum_{i=1}^{L}D_N(Nq_{1i},Nq_{2i})\right)}{\sum_{Q}P(Q)} & \text{if } p_1 \neq p_2\\ 0 & \text{if } p_1 = p_2\end{cases} \qquad (1)$$

where Q is an alignment between the states of the HMMs of the phones p_1 and p_2, $P(Q)$ is the probability of Q, L is the length of the alignment, q_{1i} and q_{2i} are states of the models that are aligned according to Q, $N_{q_{1i}}$ and $N_{q_{2i}}$ are the Gaussian distributions associated to the states q_{1i} and q_{2i}, and $D_N(\cdot)$ is a measure of distance between the two Gaussian distributions. The numerator is a weighted sum of the average distance between the Gaussians of the aligned states for each alignment Q. Bahlmann and Burkhardt (2001) calculated this average distance between Gaussians for each Q and chose the minimum one. On the other hand, we sum all these average Gaussian distances weighted by the probability of the alignment. Since only a subset of the possible alignments is used, the denominator is introduced in order to normalise by the probability of the subset of alignments. In this work, we used the alignments associated to the possible paths in a grid of dimension M_1xM_2, where M_1 and M_2 are the number of states of the models as shown in Fig. 1. This subset avoids alignments where there are loops in states of the two models at the same time.

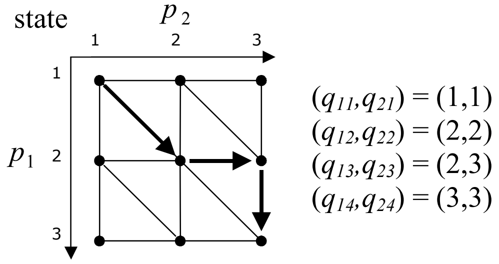
state
p_2
1
2
3
1
2
3
p_1
$(q_{11},q_{21}) = (1,1)$
$(q_{12},q_{22}) = (2,2)$
$(q_{13},q_{23}) = (2,3)$
$(q_{14},q_{24}) = (3,3)$

Fig. 1 Subset of alignments used to calculate the inter-HMM distance. The bold line shows one of these alignments. The values of q_{1i} and q_{2i} are the aligned states according to the path in bold.

The models used to obtain a dissimilarity value between the phones with the proposed measure have one Gaussian per state. This does not imply that the real ASR systems must have one Gaussian per state. We considered several monomodal Gaussian distances such as Euclidean, Mahalanobis, Kullback-Leibler, Bhattacharyya and Jeffreys-Matusita [4,5]:

Euclidean distance:

$$D_{EUC}(N_1,N_2)=(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) \tag{2}$$

Bhattacharyya distance:

$$\begin{aligned}D_{BHA}(N_1,N_2)=&\frac{1}{8}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\left[\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right]^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\\&+\frac{1}{2}\log\frac{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\end{aligned} \tag{3}$$

Jeffreys-Matusita distance:

$$D_{JM}(N_1,N_2)=\sqrt{2}(1-\exp(D_{BHA}(N_1,N_2)))^{1/2} \tag{4}$$

Kullback-Leibler distance:

$$\begin{aligned}D_{KL}(N_1,N_2)=&\frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\left[\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right](\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\\&+\frac{1}{2}tr\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1-2\mathbf{I}\right)\end{aligned} \tag{5}$$

Mahalanobis distance:

$$D_{MAH}(N_1,N_2)=(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\left[\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right]^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) \tag{6}$$

where $\boldsymbol{\mu}_i$ and Σ_i are the mean vector and the covariance matrix of the Gaussian N_i respectively.

2.2 Inter-Phone Distance Based on Phonetic Knowledge

Another way to obtain a distance between two phones is to use the knowledge of their phonetic characteristics (Covington, 1996; Somers, 1998; Kondrak, 2000).

Before the definition of this distance, we have divided the phones into groups according to their characteristics. In this study we have worked with French words, and in this language the main groups are:

$$g(p)=\begin{cases}Vowel(V)\\Glide\\Liquid\\Fricative\\Stop\\Nasal\ Consonant\end{cases} \tag{7}$$

where *g(p)* is the group the phone *p* belongs to. With this classification of the phones, we have defined three different inter-phone distances based on phonetic knowledge:

$$d_{PK}^{(1)}(p_1,p_2)=\begin{cases}0 & \text{if } p_1=p_2\\ \alpha & \text{if } p_1\neq p_2\end{cases} \tag{8}$$

$$d_{PK}^{(2)}(p_1,p_2)=\begin{cases}0 & \text{if } (p_1=p_2)\\ \gamma & \text{if } (p_1\neq p_2)\,\&\,(g(p_1)=g(p_2))\\ \alpha & \text{if } (p_1\neq p_2)\,\&\,(g(p_1)\neq g(p_2))\end{cases} \tag{9}$$

$$d_{PK}^{(3)}(p_1,p_2)=\begin{cases}0 & \text{if } (\mathrm{p}_1=\mathrm{p}_2)\,\&\,(g(p_1)=g(p_2)=V)\\ \sigma & \text{if } (\mathrm{p}_1\neq\mathrm{p}_2)\,\&\,(g(p_1)=g(p_2)=V)\\ \gamma & \text{if } (\mathrm{p}_1=\mathrm{p}_2)\,\&\,((g(p_1)=g(p_2))\neq V)\\ \beta & \text{if } (\mathrm{p}_1\neq\mathrm{p}_2)\,\&\,((g(p_1)=g(p_2))\neq V)\\ \alpha & \text{if } (\mathrm{p}_1\neq\mathrm{p}_2)\,\&\,(g(p_1)\neq g(p_2))\end{cases} \tag{10}$$

where 0<σ<γ<β<α are constant values.

The distance $d^{(1)}_{PK}(p_1,p_2)$ is the simplest one and gives a high distance if two phones are different and 0 if they are equal. The distance $d^{(2)}_{PK}(p_1,p_2)$ is similar but gives a medium distance if the phones are different but belong to the same group. On the other hand the distance $d^{(3)}_{PK}(p_1,p_2)$ gives low distances if both phones are vowels and higher distances if at least one of the phones is not a vowel. As it will be explained in section 4, these distances are used to align phonetic transcriptions. Therefore, different alignments are obtained depending on the used distance.

3 Alignment between Phonetic Transcriptions

Once we have an inter-phone distance, in order to calculate a distance between the phonetic transcriptions of two words, we first need to align them. Let $W_1=\{p_{1i}\}$ and $W_2=\{p_{2j}\}$, with i=1,…,I and j=1,...,J, be the phonetic transcriptions of the two words to compare. The values I and J are the lengths of the phonetic transcriptions and p_{1i} and p_{2j} are their phones. Let us consider an *i-j* grid, shown in Fig. 2, where W_1 and W_2 are developed along the *i*-axis and the *j*-axis respectively. A path through the grid is written as $F=\{c(1),c(2)...c(K)\}$, and it represents an alignment between the two transcriptions. The generalised element of the path is *c(k)* and it consists of a pair of coordinates in the *i* and *j* directions. The *i* and *j* coordinates of the *k*th path element are *i(k)* and *j(k)* respectively.

$$c(k)=(i(k),j(k)) \tag{11}$$

The path F fulfils the following conditions (Sakoe and Chiba, 1978):

1) Monotonic conditions:

$$i(k-1) \le i(k) \text{ and } j(k-1) \le j(k) \tag{12}$$

2) Continuity conditions:

$$i(k) - i(k-1) \le 1 \text{ and } \mathrm{j}(k) - j(k-1) \le 1 \tag{13}$$

3) Boundary conditions

$$i(K) = I \text{ and } j(K) = J \tag{14}$$

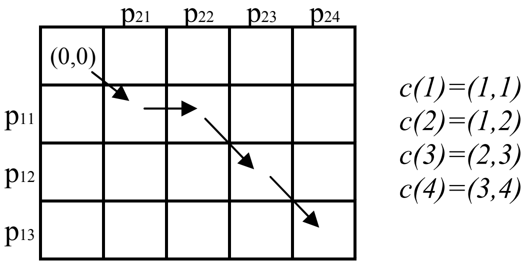

Fig 2. Example of a path F in the grid, and the steps *c(k)*. Each path defines an alignment between the phonetic transcriptions.

We have used two different kinds of alignments. How an alignment is defined by the path F is different for each one. These kinds of alignments are described in the following sections.

3.1 Alignment with Only Substitutions

This is the classical alignment used in DTW (Sakoe and Chiba, 1978), where only substitutions are allowed. We denote this kind of alignment as OS (Only Substitutions). When using this alignment each element *c(k)* indicates that the phones $p_{1i(k)}$ and $p_{2j(k)}$ are aligned. For example the OS alignment defined with the path of Fig. 2 is the following one:

p_{21}	p_{22}	p_{23}	p_{24}
p_{11}	p_{11}	p_{12}	p_{13}

The OS alignment is the one used in previous works (Tan et al., 1999) and it implies that one phone of a phonetic transcription can be aligned with more than one phone of the other phonetic transcription. But phones of the same transcription are always different. This is the reason why we also use the ID alignment, explained in the following section.

3.2 Alignment with Substitutions Insertions and Deletions

This kind of alignment is usually used in the alignment of phonetic or DNA sequences (Waterman and Eggert, 1987), and it allows insertions and deletions. We denote it as ID alignment. When using this kind of alignment the alignment is defined by the path F as follows:

- if *i(k)=i(k-1)+1* and *j(k)=j(k-1)+1* then $p_{1i(k)}$ and $p_{2j(k)}$ are aligned.
- if *i(k)=i(k-1)+1* and *j(k)=j(k-1)* then $p_{1i(k)}$ is aligned with the null character (symbol of an insertion or an omission)
- if *i(k)=i(k-1)* and *j(k)=j(k-1)+1* then $p_{2j(k)}$ is aligned with the null character.

The ID alignment defined with the path of Fig. 2 is the following one:

p_{21}	p_{22}	p_{23}	p_{24}
p_{11}	–	p_{12}	p_{13}

If the ID alignment is used the inter-phone distances have to be extended to cover pairs consisting of a phone and the null character, which corresponds to the operation of insertion or deletion. The inter-phone distance calculated from the acoustic models for the ID alignment is as follows:

$$d_{ID-HMM}(c(k)) = \begin{cases} d_{HMM\_} & \text{if} \begin{pmatrix} i(k) = i(k-1) \text{or} \\ j(k) = j(k-1) \end{pmatrix} \\ d_{HMM}(p_{1i(k)}, p_{2j(k)}) & \text{otherwise} \end{cases} \tag{15}$$

where $d_{HMM\_}$ is the distance between a phone and the null character. We have chosen the following value:

$$d_{HMM\_} = \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} d_{HMM}(p_i, p_j) \tag{16}$$

where P is the total number of phones. The value of the equation (16) is the average of the distance between all the phones.

We have to do the same with the inter-phone distance based on phonetic knowledge:

$$d_{ID-PK}^{(n)}(c(k)) = \begin{cases} d_{PK\_} & \text{if} \begin{pmatrix} i(k) = i(k-1) \text{or} \\ j(k) = j(k-1) \end{pmatrix} \\ d_{PK}^{(n)}(p_{1i(k)}, p_{2j(k)}) & \textit{otherwise} \end{cases} \tag{17}$$

where $d_{PK\_} > \alpha$ is the distance between a phone and the null character, and n can be 1, 2 or 3, any of the distances of the equations (8), (9) or (10).

4 *Inter-Word Distances*

The proposed application of this work is to predict if two words are likely to be confused by an ASR system, i.e, if they are confusable or not. In order to do this, a distance is calculated between the two words and, if the distance is lower than a threshold, the word pair is considered confusable:

$$\begin{cases} \text{if } D_*(W_1,W_2) \le \textit{Threshold} \Rightarrow \textit{Confusable} \\ \text{if } D_*(W_1,W_2) > \textit{Threshold} \Rightarrow \textit{Not Confusable} \end{cases}$$

where $D_*(W_1,W_2)$ is a distance between two words, that can be any one of the proposed in the following two sections.

4.1 Dynamic Time Warping

Based on the alignments definitions and the inter-phone distances presented in the previous sections, a distance between two words is defined as follows (Sakoe and Chiba, 1978):

$$D_{OS-DTW}(W_1,W_2) = \min_F \left[\frac{\sum_{k=1}^{K} d_{HMM}\left(p_{1i(k)}, p_{2j(k)}\right) w(k)}{\sum_{k=1}^{K} w(k)} \right] \quad (18)$$

The weighting function $w(k)$ introduced into the overall distance measure is used to normalise for the path length. In this work we have used the following weighting function (Sakoe and Chiba, 1978):

$$w(k) = i(k) - i(k-1) + j(k) - j(k-1) \quad (19)$$

This implies that:

$$\sum_{k=1}^{K} w(k) = I + J \quad (20)$$

With this weighting function the denominator of equation (18) is constant and, therefore, independent of the path F.

The equation (18) means that the distance between two words is the minimum weighted summation of the distances between the aligned phones, for all the possible OS alignments between the phonetic transcriptions of the words

This distance is based on the OS alignment defined in section 3. If we want to use the ID alignment we only have to replace the distance $d_{HMM}(p_1,p_2)$ with $d_{ID-HMM}(c(k))$ in (18).

We call this distance DTW, OS-DTW or ID-DTW depending on the used alignment, because it is usually used in the Dynamic Time Warping technique.

4.2 Phonetic Acoustic Dissimilarity Measure

The DTW technique searches the alignment that minimizes the accumulated distance. This may cause two words that are not confusable to have a low dissimilarity value when it should be high. For this reason in this paper we propose a modification of DTW that we call Phonetic Acoustic Dissimilarity measure (PAD). The difference between them is that, when using the PAD measure, the alignment is based on phonetic information, not in acoustic information. The acoustic information is only used to calculate the accumulated distance once the alignment is done. The PAD measure with the OS alignment is calculated as follows:

$$D_{OS-PAD}(W_1,W_2) = \frac{\sum_{k=1}^{K} d_{HMM}\left(p_{1i^*(k)}, p_{2j^*(k)}\right) w(k)}{\sum_{k=1}^{K} w(k)} \quad (21)$$

where $i^*(k)$ and $j^*(k)$ are the coordinates of the alignment $F^* = \{c^*(1), c^*(2) \ldots c^*(K)\}$. This alignment is:

$$F^* = \arg\min_F \left[\frac{\sum_{k=1}^{K} d_{PK}^{(n)}\left(p_{1i(k)}, p_{2j(k)}\right) w(k)}{\sum_{k=1}^{K} w(k)} \right] \quad (22)$$

and n can be 1, 2 or 3, any of the distances of the equations (8), (9) or (10). This alignment can be efficiently found by Dynamic Programming. We can see that here, the inter-phone distance used to obtain the alignment between the phonetic transcriptions is not the same that the one used to calculate the inter-word distance. The first one is based on phonetic knowledge, and the second one is obtained from the acoustic models of the phones, as we have explained in section 2. With this modification we can define $d^{(n)}_{PK}(c(k))$ so that we use an alignment that we consider correct to calculate the inter-word distance, rather than the alignment that gives the lower inter-word distance as in DTW.

The distance of the equation (21) is based on the OS alignment. We can use the ID alignment instead of the OS alignment with the PAD measure in the same way that we have done with the DTW distance. We only have to replace $d_{HMM}(p_1,p_2)$ with $d_{ID-HMM}(c(k))$ in equation (21), and $d^{(n)}_{PK}(p_1,p_2)$ with

$d^{(n)}_{ID\text{-}PK}(c(k))$ in the equation (22), and we obtain the ID-PAD distance.

5 *Experiments and Results*

5.1 Experimental Setup

In order to evaluate the performances of this method, we introduce a classical false acceptance/false rejection framework for comparing a posteriori classification obtained by testing ASR systems with the a priori classification produced by the method. We constructed two kinds of ASR systems: one to detect the confusable word pairs, and the other to detect the not confusable word pairs:

NCD Systems (No Confusability Detection): 223 systems, each one with only one word in its vocabulary and a garbage model to reject out-of-vocabulary data. Each system has been tested with the 223 words.

CD System (Confusability Detection): One system with 841 words and a garbage model, tested with the 841 words.

If one of the NCD systems, with only the word A in its vocabulary, is tested with another word B and they are never confused, it means that they are very different and, therefore, they are not confusable. On the other hand, if they are sometimes confused, it only means that B is more similar to A than to the garbage model, not necessarily that A and B are similar. Therefore, with this kind of systems we can only determine if two words are not confusable in general.

If we test the CD system with several utterances of a word A, and a word B is never recognized, we cannot say that A and B are not confusable, we can only say that A is more similar to some of the other words of the vocabulary than to B. On the other hand, if they are sometimes confused, we can assure that they are quite confusable. Therefore, with this system we can detect confusable word pairs.

The vocabulary of the CD and NCD systems consisted of French isolated words such as numbers, cities, commands, etc. Each word was pronounced by 700 speakers in average. The speech signal was sampled at 8 kHz and parameterized using MFCCs. The feature vectors consisted of 27 coefficients: the frame energy, 8 MFCCs, and the first and second time derivatives. The models of the words were constructed by concatenating context dependent HMMs of the phones with one Gaussian per state. By testing these systems the following three groups of word pairs are obtained:

Low Probability of Confusion (LPC): 21506 word pairs which were never confused when the NCD systems were tested.

Medium Probability of Confusion (MPC): 150 word pairs which had a confusion rate lower than 5% and higher than 0% when the CD system was tested.

High Probability of Confusion (HPC): 189 word pairs which had a confusion rate higher than 5% when the CD system was tested.

It would have been desirable to have more HPC word pairs to better rely on the results, but no more words were available. We consider a False Rejection to classify as confusable an LPC word pair, and a False Acceptance to classify as not confusable a HPC word pair. The MPC word pairs were not taken into account in the evaluation because we considered that is not a severe error neither to classify them as confusable nor as not confusable.

We used the following values (8), (9), (10) and (17): α=4, β=3, γ=2, σ=1 and $d_{PK\_}$=7. The HMMs used to calculate the inter-phone distances are not the models used in recognition. In the first case we used models without context with 3 states and 1 Gaussian per state. The results would probably improve by using context dependent models, but the complexity of the system would increase.

5.2 Confusability Prediction Results

Table 1 shows the EER for each inter-word distance, each Gaussian distance and the OS alignment. The EER is the False Acceptance Rate and the False Rejection Rate obtained with the threshold that makes them equal. We denote as PAD1, PAD2 and PAD3, the PAD measures calculated with $d^{(1)}_{PK}(c(k))$, $d^{(2)}_{PK}(c(k))$ and $d^{(3)}_{PK}(c(k))$ respectively. We can see that the proposed PAD measure always outperforms the classical DTW distance, independently of the inter-phone distance used to do the alignment. We can also see that OS-PAD3 gives lower error rates than OS-PAD2, and that OS-PAD2 gives lower error rates than OS-PAD1. Therefore, for our purpose, it is better to give priority to align vowels rather than to align other phones, because this is the basis of the OS-PAD3 measure.

	OS-DTW	OS-PAD1	OS-PAD2	OS-PAD3
EUC	9,4%	7,9%	6,9%	6,8%
KL	9,6%	9,0%	7,9%	6,9%
JM	11,7%	9,5%	9,4%	9,0%
MAH	12,1%	9,8%	8,5%	7,9%
BHA	17,0%	16,0%	14,6%	13,8%

Table 1. The EER for each inter-word distance with the OS alignment and each Gaussian distance in equation (1).

Table 1 also shows that, with the OS alignment, the best Gaussian distance is the Euclidean, because it gives the lower error rates independently of the used inter-word distance. With this Gaussian distance a 6.8% of EER is obtained when using OS-PAD3.

	ID-DTW	ID-PAD1	ID-PAD2	ID-PAD3
EUC	3,1%	2,1%	2,1%	2,1%
KL	3,2%	1,6%	1,6%	1,6%
JM	7,5%	6,9%	6,5%	6,3%
MAH	2,6%	2,6%	2,6%	2,5%
BHA	8,9%	10,1%	9,6%	8,9%

Table 2. The EER for each inter-word distance with the ID alignment and each Gaussian distance in equation (1).

Table 2 shows the same results as in table 1 but with the ID alignment instead of OS. The first conclusion obtained when comparing the table 2 with the table 1 is that the ID alignment gives better results. The ID alignment always outperforms the OS alignment independently of the inter-word distance and the Gaussian distance. ID-PAD outperforms ID-DTW for all the Gaussian distances except Bhattacharyya. ID-PAD3 outperforms ID-PAD2, and ID-PAD2 outperforms ID-PAD1, except with the Euclidean and Kullback-Leibler distances. In these cases the three PAD measures give the same results. The best results are obtained with the Kullback-Leibler Gaussian distance, with a 1.6% of EER. The Euclidean distance also gives low error rates and it has lower computational cost.

In Fig. 3 we can see the FAR and FRR curves for the ID-PAD3 and ID-DTW distances with the KL Gaussian distance. We can see that the FAR curve is similar for the two inter-word distances. This implies that they do a similar alignment when the words to compare are similar. On the other hand, the FRR curve of the ID-PAD3 distance is lower than that of the ID-DTW distance. This implies that ID-PAD3 gives higher distances to the word pairs that are different, making a better separation between the two classes, confusable and not confusable.

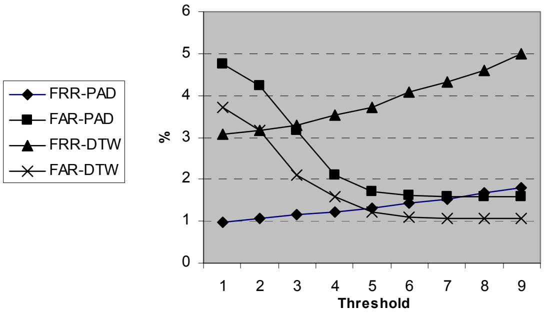

Fig. 3. FAR and FRR curves for the ID-PAD3 and ID-DTW distances, with the KL Gaussian distance.

6 *Conclusions*

In this paper we have proposed a new inter-word distance based on DTW called PAD. This new distance is based on an alignment that is obtained with an inter-phone distance that is not the same that the one used to calculate the inter-word distance. The first one is based on phonetic knowledge, and the second one is obtained from the acoustic models of the phones. We have proposed new inter-phone distances of both types. We have also used two different alignments: either with or without insertions and deletions.

We have applied the new inter-word distance to predict if two words are likely confused by and ASR system. The new distance outperformed de classical DTW in terms of EER, with a 50% of EER reduction in the best case. The alignment with insertion and deletions provided lower error rates. Using this alignment and the PAD measure an EER of 1.6% is obtained.

References

Claus Bahlmann and Hans Burkhardt. “Measuring HMM similarity with the Bayes probability of error and its application to online handwriting recognition”.

Proceedings of the ICDAR, pp. 406-411, 2001

Michèle Basseville. 1989. Distance measures for signal processing and patter recognition. *Signal Processing*, Vol. 18(4), pp. 349-369.

Michael A. Covington. 1996. An algorithm to align words for historical comparison. In *Computational Linguistics,*22(4):481-496.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*.

Sandrine Pouységur. 2001. Etude du taux de confusion de mots pour la reconnaissance de mots isolés. *4e Rencontres jeunes chercheurs en parole*.

Lawrence Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proceedings of the IEEE*, 77(2):257-286.

David B. Roe and Michael D. Riley. 1994. Prediction of word confusabilities for speech recognition. In *Proceedings of the ICSLP*, pp. 227-230.

Hiroaki Sakoe and, Seibi Chiba. 1978. Dynamic Programming algorithm optimization for spoken word recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. ASSP-26, N°1.

Harold L. Somers. 1998. Similarity metrics for aligning children's articulation data. In *Proceedings of COLING-ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp 1227-1231

Harold L. Somers. 1999. Aligning phonetic segments for children's articulation assessment. *Computational Linguistics, 25(2):267-275.*

Jayren J. Sooful and Elizabeth C. Botha. 2001. An acoustic distance measure for automatic cross-language phoneme mapping. In *Proceedings of the PRASA*.

Beng T. Tan, Yong Gu and Trevor Thomas. 1999. Word Confusability Measures for Vocabulary Selection in Speech Recognition. In *Proceedings of the ASRU.*

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery.*

Michael S. Waterman and Mark Eggert. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.*, 197:723-728.