Word Confusability Prediction in Automatic Speech Recognition

Jan Anguita[*], *Stéphane Peillon*[+], *Javier Hernando*[*], *Alexandre Bramoullé*[+]

[*]TALP Research Center, Universitat Politècnica de Catalunya, Dept. of Signal Theory and Communications, Barcelona, Spain.
[+]Telisma R&D, Lannion, France.
{jan,javier}@talp.upc.es
speillon@telisma.com

Abstract[1]

A new method to predict if two words are likely to be confused by an Automatic Speech Recognition (ASR) system is presented in this paper. A new inter-word dissimilarity measure based on Dynamic Time Warping (DTW) is used to classify the word pairs as confusable or not confusable. Firstly, the phonetic transcriptions of the two words to compare are aligned using only phonetic information. After the alignment, the accumulated distance is obtained with a new inter-phone acoustic distance calculated between the Hidden Markov Models (HMM) of the phones. In addition, we have used two different kinds of alignment: either with or without insertions and omissions. 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. The prediction Equal Error Rate (EER) was measured to be 1.6%, a 50% of reduction with respect to the conventional DTW distance.

1. Introduction

Confusion errors can be a serious problem in Automatic Speech Recognition (ASR) systems. For example, in an application involving commands, the system will not do the action that the user ordered if confusion occurs. Therefore, if possible, the vocabulary of a speech recognition application should be chosen to avoid confusable words. In [1,2,3] there are some proposals to measure the confusability between words in order to help to design the vocabulary of an ASR system. Working on the principle that phonetically similar words are more easily confused by ASR systems, they propose inter-word dissimilarity measures in order to choose the words of the vocabulary to be as less similar as possible. In [1], a dissimilarity measure between words is used to choose the least confusable words from a list.

In this work we propose to take a decision after the dissimilarity measure and classify the word pairs in two classes: confusable or not confusable, i.e., predict if two words are likely to be confused by ASR systems or not. This approach provides a powerful tool that can be used to avoid confusable words in the vocabulary.

We also propose a new inter-word dissimilarity measure that is based on Dynamic Time Warping (DTW) [4]. This new measure, which we call Phonetic Acoustic Dissimilarity (PAD) measure, is calculated in two steps. Firstly, the phonetic transcriptions of the two words are aligned using a Dynamic Programming algorithm where the local distances are obtained from the phonetic characteristics of the phones. Secondly, the accumulated distance is calculated on the basis of the resulting alignment. In the second step we use a local distance that is calculated with a new distance between the Hidden Markov Models (HMM) of the phones, which carry acoustic information.

In order to test the method, a classical false acceptance/false rejection framework is introduced. We propose a procedure to determine which words are usually confused by ASR systems to compare it with the classification of our method.

In addition, we have implemented both the DTW and the PAD measures with two different kinds of alignment: either with or without insertions and omissions. Our results show that with the alignment with insertions and omissions lower classification errors are obtained, especially with PAD.

In section 2 two kinds of inter-phone distances are proposed. The first one is based on phonetic knowledge. The second one is obtained from the HMMs of the phones. Section 3 describes the two alignments used in this work: either with or without insertions and omissions. In section 4 the conventional DTW is reviewed and the new PAD measure is presented. Section 5 describes how to obtain the data to test, the performed experiments and the obtained results. Finally, section 6 contains the conclusions of this work.

2. Inter-Phone Distance Measures

2.1. Distances based on Phonetic Knowledge

One way to obtain a distance between two phones is to use the knowledge of their phonetic characteristics [5]. In this paper we propose the following inter-phone distances:

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

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

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

[1] This paper reports the work performed during an internship at Telisma (France).

where $0<\sigma<\gamma<\beta<\alpha$ are constant values, p_1 and p_2 are phones, and $g(p)$ is the group the phone p belongs to. We divided the phones into the following groups: *Vowel(V), Glide, Liquid, Fricative, Stop* and *Nasal consonant*.

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.

2.2. Distance Measure between Hidden Markov Models

Another way to obtain a distance measure between two phones is to calculate the distance between their acoustic models. Since in modern ASR systems the acoustic units are usually modeled by HMMs, in this paper we propose the following distance measure between two HMMs:

$$d_{HMM}(p_1,p_2)=\begin{cases}\dfrac{\sum_{Q}P(Q)\dfrac{1}{L}\sum_{i=1}^{L}D_N\left(N_{q_{1i}},N_{q_{2i}}\right)}{\sum_{Q}P(Q)} & \text{if } p_1\neq p_2\\ 0 & \text{if } p_1=p_2\end{cases} \quad (4)$$

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. In [6], this average distance between Gaussians is calculated for each Q and the minimum one is chosen. 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 have used the alignments associated to the possible paths in a grid of dimension $M_1 \times M_2$, where M_1 and M_2 are the number of states of the models. This subset avoids alignments where there are loops in states of the two models at the same time.

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 and Kullback-Leibler [7,8].

3. Phonetic Alignments

Let $W_1=\{p_{1i}\}$ and $W_2=\{p_{2j}\}$, with $i=1,\ldots,I$ and $j=1,\ldots,J$, 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 $I \times J$ grid (Fig. 1), where W_1 and W_2 are placed along the i-axis and the j-axis respectively. A path through the grid is denoted as $F=\{c(1),c(2)\ldots c(K)\}$, and it represents an alignment between the two transcriptions. The path fulfills the monotonic and continuity conditions described in [4]. Each element of the path $c(k)$ consists of a pair of coordinates $(i(k),j(k))$ that indicate a point in the grid.

In this work we have considered two kinds of alignment that we denote as OS (Only Substitutions) and IO (Insertions and Omissions) respectively. When using the OS alignment only substitutions are allowed, i.e., each element $c(k)$ indicates that the phones $p_{1i(k)}$ and $p_{2j(k)}$ are aligned. On the other hand, when using the IO alignment not only substitutions are permitted but also insertions and omissions. In this case, 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.

Fig. 1 shows an example of how a path F defines an OS or an IO alignment.

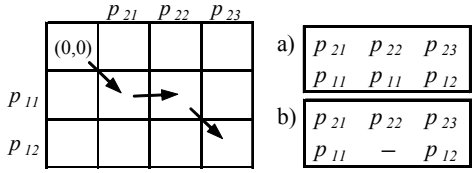

Figure 1: Example of a path F and its associated OS (a) and IO (b) alignments.

4. Measures for Word Confusability Prediction

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. For this purpose, a distance is calculated between the two words and, if it is lower than a threshold, the word pair is considered confusable:

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

where $D_*(W_1,W_2)$ is a distance between two words. In the following sections several proposals are presented.

4.1. Dynamic Time Warping

The conventional Dynamic Time Warping (DTW) technique [4] can be used to calculate a distance between two phonetic transcriptions [1]. Based on the OS alignment and the HMM-based inter-phone distance presented in the previous sections, the DTW distance between two words is defined as follows:

$$D_{OS-DTW}(W_1,W_2)=\min_{F}\left[\frac{\sum_{k=1}^{K}d_{HMM}\left(p_{1i(k)},p_{2j(k)}\right)w(k)}{\sum_{k=1}^{K}w(k)}\right] \quad (5)$$

where $d_{HMM}(p_{1i(k)}, p_{2j(k)})$ is the distance between HMMs presented in section 2 and $w(k)$ is a weighting function introduced to normalise by the path length. In this work we have used the following one [4]:

$$w(k) = i(k) - i(k-1) + j(k) - j(k-1) \quad (6)$$

The OS-DTW distance 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.

In order to use the IO alignment, the inter-phone distance has to be extended to cover pairs consisting of a phone and the null character, which corresponds to the operation of insertion or omission:

$$d_{IO-HMM}(c(k)) = \begin{cases} d_{HMM\_} & \text{if} \begin{pmatrix} i(k) = i(k-1) \text{ or} \\ j(k) = j(k-1) \end{pmatrix} \\ d_{HMM}\left(p_{1i(k)}, p_{2j(k)}\right) & \text{otherwise} \end{cases} \quad (7)$$

where $d_{HMM\_}$ is the distance between a phone and the null character. This value was set at the arithmetic mean of the distances between all the phones:

$$d_{HMM\_} = \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} d_{HMM}\left(p_i, p_j\right) \quad (8)$$

where P is the total number of phones. The DTW distance with the IO alignment (IO-DTW distance) is obtained by replacing $d_{HMM}(p_{1i(k)}, p_{2j(k)})$ by $d_{IO-HMM}(c(k))$ in (5).

4.2. Phonetic Acoustic Dissimilarity Measure

The DTW technique forces the alignment that minimizes the accumulated distance, which may cause classification errors. For this reason, in this paper we propose a modification of the DTW distance that we call Phonetic Acoustic Dissimilarity (PAD) measure. 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:

$$D_{OS-PAD}(W_1, W_2) = \frac{\sum_{k=1}^{K} d_{HMM}\left(p_{1i^*(k)}, p_{2j^*(k)}\right) w(k)}{\sum_{k=1}^{K} w(k)} \quad (9)$$

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

$$F^* = \arg\min_{F} \left[\frac{\sum_{k=1}^{K} d_{PK}^{(n)}\left(p_{1i(k)}, p_{2j(k)}\right) w(k)}{\sum_{k=1}^{K} w(k)} \right] \quad (10)$$

where n can be 1, 2 or 3 and $d^{(n)}_{PK}(p_{1i(k)}, p_{2j(k)})$ is one of the inter-phone distances presented in section 2.1. We can see that firstly, an alignment between the phonetic transcriptions of the words is done based on the $d^{(n)}_{PK}(p_{1i(k)}, p_{2j(k)})$ inter-phone distance. After, the inter-word distance is calculated with the resulting alignment and the $d_{HMM}(p_{1i(k)}, p_{2j(k)})$ inter-phone distance. Depending on the chosen value of n we obtain three different distances: OS-PAD1, OS-PAD2 or OS-PAD3. The difference between them is the alignment obtained with (10). With $d^{(1)}_{PK}(c(k))$ we give priority to align equal phones, with $d^{(2)}_{PK}(c(k))$ to align equal phones and phones of the same group, and with $d^{(3)}_{PK}(c(k))$ to align vowels.

In case of using the IO alignment, the $d^{(n)}_{PK}(p_{1i(k)}, p_{2j(k)})$ inter-phone distance has to be extended to cover pairs consisting of a phone and the null character to obtain the $d^{(n)}_{IO\text{-}PK}(c(k))$ distance (as we did with the $d_{HMM}(p_{1i(k)}, p_{2j(k)})$ distance in (7)). In this case, the distance between a phone and the null character is a constant value $d_{PK\_}$. The PAD measure with the IO alignment (IO-PAD measure) is obtained by replacing $d_{HMM}(p_{1i(k)}, p_{2j(k)})$ by $d_{IO-HMM}(c(k))$ in (9), and $d^{(n)}_{PK}(p_{1i(k)}, p_{2j(k)})$ by $d^{(n)}_{IO-PK}(c(k))$ in (10). Depending on the chosen value of n we obtain three different distances: IO-PAD1, IO-PAD2 or IO-PAD3.

5. Experiments and Results

5.1. Experimental Setup

In order to test our method we need to determine which pairs of words are usually confused by ASR systems to compare them with the prediction of our method. We constructed two kinds of ASR systems: one to detect confusable word pairs, and the other to detect not confusable word pairs.

- **DNC Systems (Detection of No Confusable words):** 223 systems, each one with only one word in its vocabulary and a garbage model to reject out-of-vocabulary data. Each system was tested with the 223 words.
- **DC System (Detection of Confusable words):** One system with 841 words and a garbage model, tested with the 841 words.

If one of the DNC 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 DC system with several pronunciations 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 DC and DNC 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 were obtained:

- **Low Probability of Confusion (LPC):** 21506 word pairs which were never confused when the DNC 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 DC system was tested.

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

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: 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.

5.2. Confusability Prediction Results

Table 1 shows the Equal Error Rate (EER) for each inter-word distance, each Gaussian distance and the OS alignment. The EER is the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) obtained with the threshold that makes them equal. We can see that the proposed PAD measure always outperforms the classical DTW distance, independently of the Gaussian distance. The OS-PAD3 measure outperforms OS-PAD2 and OS-PAD1 in all cases. The lowest EER, 6,8%, is obtained with the OS-PAD3 inter-word distance and the Euclidean Gaussian distance.

	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%
MAH	12,1%	9,8%	8,5%	7,9%

Table 1: EER obtained with the OS-DTW and OS-PAD measures for each Gaussian distance in (4)

Table 2 shows the same results that Table 1 but with the IO alignment instead of OS. The first conclusion obtained when comparing Table 2 with Table 1 is that the IO alignment outperforms the OS alignment independently of the inter-word and the Gaussian distances. With the IO alignment there is almost no difference between IO-PAD1, IO-PAD2 and IO-PAD3. The lowest EER, 1.6%, is obtained with the Kullback-Leibler Gaussian distance, with a 50% of EER reduction from IO-DTW to IO-PAD.

	IO-DTW	**IO-PAD**	**IO-PAD2**	**IO-PAD3**
EUC	3,1%	2,1%	2,1%	2,1%
KL	3,2%	**1,6%**	**1,6%**	**1,6%**
MAH	2,6%	2,6%	2,6%	2,5%

Table 2: EER obtained with the IO-DTW and IO-PAD measures for each Gaussian distance in (4)

In Fig. 2 we can see the FAR and FRR curves for the IO-DTW and IO-PAD3 measures, 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 IO-PAD3 measure is lower than that of the IO-DTW distance. This implies that IO-PAD3 gives higher distances to the word pairs that are different, making a better separation between the two classes, confusable and not confusable.

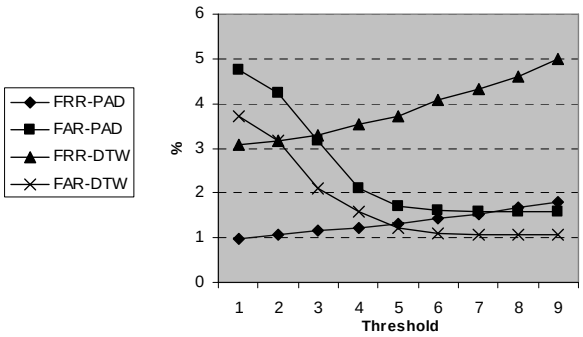

Figure 2: The FAR and FRR curves of the IO-DTW and IO-PAD3 measures with the KL Gaussian distance.

6. Conclusions

In this paper we have proposed a dissimilarity measure between phonetic transcriptions and a distance between HMMs that we use to detect confusable word pairs. A method to obtain the data to test has also been proposed.

The proposed PAD measure outperformed the classical DTW distance in terms of EER. An EER of 1.6% was obtained with the new PAD measure, which only uses phonetic information to align. Our results also show that lower EERs can be obtained by using an alignment with not only substitutions, but also insertions and omissions.

7. References

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

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

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

[4] Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming algorithm optimization for spoken word recognition". *In IEEE Trans. on ASSP*, vol. ASSP-26, Nº1, 1978.

[5] Grzegorz Kondrak. "A new algorithm for the alignment of phonetic sequences". *Proc. of the NAACL,* 2000.

[6] 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.

[7] M. Basseville. "Distance Measures for Signal Processing and Patter Recognition". *Signal Processing*, Vol. 18(4), pp. 349-369, 1989.

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