Word Confidence Using Duration Models

Stefano Scanzio [1], *Pietro Laface* [1], *Daniele Colibro* [2], *Roberto Gemello* [2]

[1] Politecnico di Torino, Italy, [2] Loquendo, Torino, Italy

{Stefano.Scanzio, Pietro.Laface}@polito.it

{Daniele.Colibro, Roberto.Gemello}@loquendo.com

Abstract

In this paper, we propose a word confidence measure based on phone durations depending on large contexts. The measure is based on the expected duration of each recognized phone in a word. In the approach here proposed the duration of each phone is in principle context-dependent, and the measure is a function of the distance between the observed and expected phone duration distributions within a word. Our experiments show that, since the “duration confidence” does not make use of any acoustic information, its Equal Error Rate (EER) in terms of False Accept and False Rejection rates is not as good as the one obtained by using the more informed acoustic confidence measure. However, combining the two measures by a simple linear interpolation, the system EER improves by 6% to 10% relative on an isolated word recognition task in several languages.

Index Terms: speech recognition, neural networks, acoustic confidence, duration confidence

1. Introduction

In any real-world application it is important to assess the reliability of each hypothesis produced by a speech recognizer. This confidence measure (CM) about the recognition results can be exploited in several different frameworks and applications: for example in Out-Of-Vocabulary (OOV) word detection, for keyword spotting, for unsupervised training, for verifying transcriptions in large corpora, for reordering the hypotheses in an N-best decoder, or even during the decoding process.

CMs are obtained by collecting during the decoding procedure some information related to acoustic as well as other features that can be useful to generate an indicator of correctness of the recognition decision.

Various proposals have been done to incorporate duration modeling in an HMM-based recognizer. Duration information can be explicitly modeled by parametric distributions of the HMM states, or indirectly by designing HMM topologies that best fit the actual phone duration distributions. Duration information can also used in the post-processing stage of speech recognition [1,2]. Confidence measures based on duration features have also been proposed in the past [3,4].

CMs obtained from phone or word duration features cannot be as accurate as acoustic based CMs, but they can be useful to reject OOV words or even In-Vocabulary recognized words with unlikely duration distributions of their component phones.

In this paper we propose the use of a duration model for computing a CM. As we will show, the “duration confidence” information combines well with the “acoustic confidence” that we take as our baseline measure. The latter is the one used, with small variations, in the LoquendoASR decoder.

We estimate the duration model for every word in the system vocabulary, using the same corpus that we use for training the acoustic models. The duration model of a word is based on the duration of its constituent phonetic units. The main idea of this work is to estimate context-dependent phone durations with a large context. Thus, the duration of each unit composing a word that frequently occurs in the training set will be actually word-dependent, whereas the duration model for a word never seen in the training set will take into account the duration of its phonetic units, each one depending on the largest context that was available in the training set.

The CM here proposed is a function of the distance between two distributions: the observed and expected phone duration distributions within a word. These durations are normalized to account for utterances of the same word pronounced with different speaking rates.

The paper is organized as follows: Section 2 gives a short overview of the LoquendoASR system, and details how its acoustic confidence measure is produced. Section 3 introduces the duration model, and the confidence measure based on the duration of the phonetic units. The experimental results are presented and commented in Section 4, and the conclusions are drawn in Section 5.

2. LoquendoASR system overview

The LoquendoASR decoder uses a hybrid HMM-ANN model where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops. The models are based on a set of vocabulary and gender independent units including stationary context-independent phones and diphone-transition coarticulation models. A Multilayer Perceptron estimates the posterior probability of each unit state, given an acoustic feature vector. The HMM transition probabilities are uniform and fixed. The ANN output layer includes a language dependent number of states (in the range 700 to 1000) [5].

2.1. Acoustic confidence measure

The acoustic confidence measure produced by the system is based on the posterior probabilities generated by the decoder.

Confidence measures based on local phone posterior probability estimates generated by a hybrid HMM/ANN model have been proposed in [6,7]. To account for the raw acoustic information associated to each frame, the best score has been proposed as a measure of the matching between the data and the model [8]. In this approach, each utterance frame is scored against every output distribution in their HMMs to find the best score, independent of any information given by the sequence of phonetic units or words.

Building on these ideas, we have proposed in [9] the Differential Confidence measure defined as:

$$DC = \frac{1}{T}\sum_{t=1}^{T}\log\left[\frac{P(s_{i^*} \mid o_t)}{\max\limits_{1\le j\le S} P(s_j \mid o_t)}\right] \qquad (1)$$

where S is the set of output states of the ANN model, o_t is the *t-th* acoustic observation vector, and s_{i*} is the sequence of states - indexed by i^* - produced by the Viterbi alignment of an utterance of T observation frames.

This score can be interpreted as the average of the confidences computed frame by frame. It produces negative values, and zero that represents maximum reliability. To be used as a confidence measure (in the range 0-1) the DC scores must be normalized. The normalization is performed by computing the state-dependent distribution of the DC score values, fitting a Gaussian, and applying the corresponding cumulative distribution function.

3. Duration based confidence

Let $U_w = \{u_1, u_2, ..., u_N\}_W$ be the sequence of stationary and transition units corresponding to the phonetic transcription of word w. Given an utterance of word w, the duration of its units $T_w = \{t_1, t_2, ..., t_N\}_W$ can be estimated, both in training and during testing, by means of Viterbi forced alignment.

The distribution of the phonetic unit durations can be estimated using the training database to obtain a duration model λ_T. The model must take into account the speaking rate, the context of the phonetic unit, and the relative duration of the units within a word.

To account for the effects of the speaking rate, the durations of the units are normalized as:

$$\hat{t}_i = \frac{t_i}{\sum_{j=1}^{N} t_j} \cdot N \qquad (2)$$

where N denotes the number of units in the word. A value of $\hat{t}_i$ greater than 1 means that the expected duration of the *i-th* unit is greater than the average unit duration within the word.

Using model λ_T and the transcription U_w of any word w, an estimation of the reference duration $R_w = \{r_1, r_2, ..., r_N\}_W$ of the units in word w can be obtained.

During testing, the distance between the test and reference durations $d(T_w, R_w)$ can be used to produce a confidence measure for the word hypothesized by the decoder.

3.1. Duration Models

Since the duration of a phone can be strongly influenced by its phonetic context, we collect the statistics of the duration of each unit separately for every context in which it appears in the training set.

In this work, experiments have been performed using three different duration models: a Regression Tree [10], an Artificial Neural Network trained to approximate the duration distributions, and a Duration Model Tree (DMT) that efficiently encodes the duration of a unit in every context frequently seen in the training data.

For the first two models the context of a given phone was fixed and limited to the previous and next 3 phones. These models are no more detailed in this paper because the size of the regression tree was too large for a real application, whereas the results of the ANN were not comparable with the performance of the DMT approach.

A DMT is a tree with a dummy root node. The nodes of the first layer are the phonetic units defined for a given language (426 for Italian). The other layers play the role of context for the nodes of the first layer: even and odd layers represent the left and right context respectively. Since by definition our transition units represent the transition between two stationary units, the context is completely defined by stationary units. Thus, the nodes of all the layers, excluding the first one, are labeled with stationary units only. Each node, with the exception of the root node, has associated the name of the unit as label and the mean duration value as attribute.

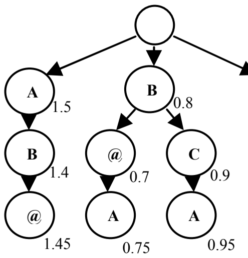

Figure 1. *Expected duration of the phonetic units of words W_1={CBA} and W_2={BA} using a small Duration Model Tree ($T(W_1)$={1.4 0.95 1.45}, $T(W_2)$= {0.75 1.45})*

The path between the root node and a given tree node encodes the context of the unit of the first layer crossed by the path. A path, thus, is associated to a given phonetic unit p. Consider a path in a DMT touching nodes labeled *A B C D E*. This path encodes phone *A* with left context (*DB*) and right context (*CE*). The mean attribute of a tree node n is the estimated duration of the unit p, in the context described by the path. The mean duration value associated with node labeled *D* in our previous example refers to phone *A* in the left context (*DB*) and right context (*C*).

Figure 1 shows an example of use of the information provided by a small DMT for obtaining the expected distribution of the duration of the units belonging to word W$_1$={CBA} and W$_2$={BA}. In the DMT, label @ refers to the silence unit. It is worth noting that the duration of pause and silence units is not modeled (these units do not appear as nodes in the first layer).

Since, in this example, unit *C* has not be seen in the training set at the beginning of a word - in the context (@)*C* - its expected duration is context-independent (1.4). The context (*C*)*B*(*A*) is present in the tree, thus the expected duration of unit *B* is 0.95, associated to node labeled *A* along the path *BCA*. The expected duration of the last unit *A* is 1.45 corresponding to the context (*B*)*A*(@).

This tree organization is memory effective because it factorizes common contexts, and allows fast search for the largest possible context of a given unit.

The duration variance can be included in the tree nodes if Gaussian or Gamma distributions are used to model the durations.

3.2. Distance between actual and expected durations

During testing, given the sequence of phonetic units of a recognized word w, and their duration, a measure of distance between the actual and the expected durations $d(T_w, R_w)$ can be used to produce a confidence measure for the word.

To account for the relative duration of the units inside a word, a scaled version of the context-dependent expected durations of the units stored in the DMT nodes is produced:

$$p_i = \frac{\hat{t}_i}{\sum_{j=1}^{N} \hat{t}_j} \qquad (3)$$

Since the parameters p_i model the duration distribution of the units inside a word, we found that the best distance measure among several used for the experiments presented in Section 4, was the Jeffries-Matusita distance, normalized by the number N of units in the word:

$$d(D_w, R_w) = \frac{\sqrt{\sum_{i=1}^{N}\left(\sqrt{p_{D,i}} - \sqrt{p_{R,i}}\right)^2}}{N}$$

(4)

The Jeffries-Matusita distance is often used in applications that require the comparison of histograms, for example image retrieval, remote sensing, or object tracking.

3.3. Duration confidence measure

Analyzing the distributions of the distances between actual unit durations and the durations obtained by using the DMT for the same set of words, it happens that short words are characterized on average by variances and distances that are larger compared with the ones obtained for long words. This happens because a distance computed on a small number of units is less precise.

Mean and variance normalization of the distance is, thus, performed computing the mean μ_l and standard deviation σ_l of the distances of words of equal length *l*.

Since in recognition the word length is known, the normalized distance is obtained as:

$$\hat{d}(D_w, R_w) = \frac{d(D_w, R_w) - \mu_l}{\sigma_l} \tag{5}$$

The final rescaling of the distance for obtaining a confidence measure is performed computing the distribution of the negative value of the distances on the training set, fitting a Gaussian, and applying the corresponding cumulative distribution function.

4. Experimental results

The quality of the duration confidence obtained with the DMT approach has been assessed using an isolated word recognition task in seven European languages. The languages are shown in Table 1, which summarizes the main features of the recognition tasks that consists in the recognition of application words collected from telephone channels. The training data for the LoquendoASR are large corpora of hundreds of hours of speech pronounced by thousand of speakers. The test data for these experiments belong to the databases collected in the SpeechDat projects [11].

Since the test data include 30 or 31 different words only, the system vocabulary has been extended to 2000 words to increase the acoustic and the duration distribution confusion among the models of the words. These additional words have been selected from the SpeechDat lists of directory assistance utterances. Moreover, the tests were also performed randomly adding car, babble, restaurant, street, etc. noise to the original signal. The range of the Signal-to-Noise Ratio of the resulting signals is 10-20 dB.

The performance of the LoquendoASR recognisers on these tasks is shown in the last column of Table 1. The average Word Accuracy (WA) is 95.8% and 80.6% in the clean and noisy tests respectively.We refer to these experiments as the In-Vocabulary tests. Another set of experiments has been done using the same vocabulary, but recognizing utterances of OOV phonetically balanced words to test the rejection capability of the duration confidence measure.

The duration models have been estimated on the same data that are used for training the acoustic models. Since the training set includes isolated utterances of the applications words, the duration model includes the entire context of the application words. The results obtained with this model represent, thus, an upper bound of the performance, possibly reachable in real applications. On the opposite side a lower bound is obtained by estimating the duration model excluding the contribution of the application words. In this case, the duration of the phonetic units will depend only on the largest context available from other words. We compare the performance of the two models to show that, although the performance of the duration model decreases in the second case, its contribution to the overall confidence is positive.

The confidence measure is computed to make an accept/reject decision based on a threshold τ. Setting a threshold introduces two kinds of errors depending on falsely accepted of falsely rejected hypotheses. The tradeoffs between the two types of errors can be appreciated by plotting a DET curve [12], where the False Reject Rate is plotted against the False Accept Rate on a Gaussian scale as a function of the threshold τ.

Figure 1 shows the DET plots for the duration confidence of the In-Vocabulary experiments, where the true and false examples necessary to create the DET plots correspond to the correctly and incorrectly recognized words respectively.

Table 1. *Features of the test databases: number of test files, and word accuracy in clean and noisy conditions.*

Database	# audio files	WA (%) Clean / Noise
Catalan	585	97.6 / 87.4
English	1357	86.3 / 62.9
French	1455	94.2 / 70.9
German	1470	98.7 / 79.7
Italian	1435	99.2 / 93.3
Portuguese	1446	96.8 / 81.8
Spanish	1455	97.5 / 88.2

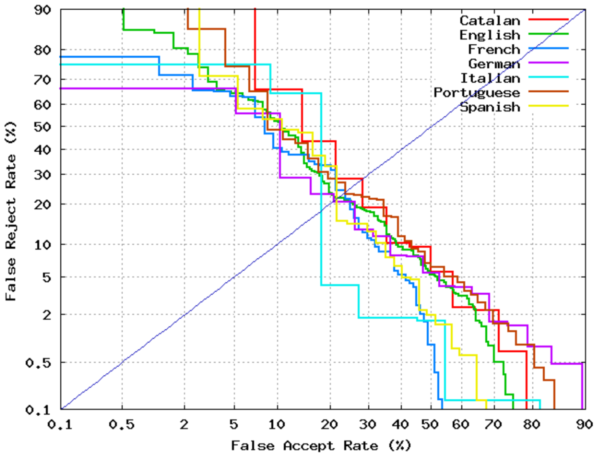

Figure 2. *DET plots for the duration confidence of the In-Vocabulary experiments.*

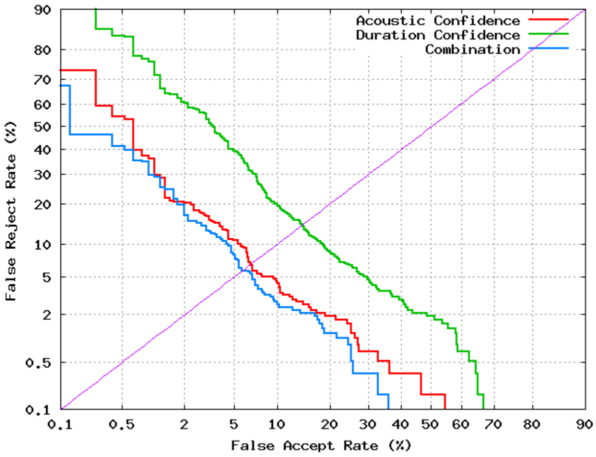

Figure 3. *DET plots of the linear combination of the acoustic and duration confidences for the Out-Of-Vocabulary experiment in Catalan language.*

Table 2. *EER for the In-Vocabulary experiments in clean / noisy condition for the acoustic and duration confidences and their combination in the best (BC) and worst case (WC).*

Database	E.E.R. (%) Acoustic	E.E.R. (%) Duration	E.E.R. (%) Fusion BC	E.E.R. (%) Fusion WC
Catalan	**14.3 / 19.8**	**28.5 / 23.1**	**14.3 / 16.7**	**14.7 / 18.8**
English	**23.1 / 26.9**	**21.9 / 28.7**	**22.6 / 24.4**	**23.4 / 25.8**
French	**16.6 / 15.3**	**23.8 / 29.7**	**16.6 / 14.4**	**16.8 / 15.0**
German	**16.7 / 19.4**	**26.3 / 33.6**	**15.8 / 20.1**	**15.8 / 20.5**
Italian	**16.8 / 11.4**	**18.2 / 18.7**	**16.1 / 10.4**	**16.5 / 11.4**
Portuguese	**19.6 / 18.6**	**23.9 / 22.8**	**19.6 / 16.8**	**19.6 / 17.5**
Spanish	**24.8 / 16.3**	**21.6 / 19.8**	**21.6 / 15.1**	**22.6 / 15.8**
Average	**18.8 / 18.2**	**23.4 / 25.2**	**18.1 / 16.8**	**18.5 / 17.8**

Table 3. *EER for the Out-Of-Vocabulary experiments in clean / noisy condition for the acoustic and duration confidences, and their combination in the best (BC) and worst case (WC).*

Database	E.E.R. (%) Acoustic	E.E.R. (%) Duration	E.E.R. (%) Fusion BC	E.E.R. (%) Fusion WC
Catalan	**6.6 / 14.6**	**14.2 / 19.3**	**5.8 / 11.7**	**6.6 / 13.1**
English	**12.0 / 19.4**	**23.0 / 27.6**	**11.2 / 17.7**	**12.5 / 19.1**
French	**6.7 / 12.7**	**22.7 / 28.9**	**6.1 / 11.9**	**6.5 / 12.7**
German	**4.0 / 12.4**	**14.7 / 24.6**	**3.5 / 11.6**	**3.9 / 11.9**
Italian	**5.6 / 12.8**	**11.2 / 17.4**	**5.1 / 11.2**	**5.3 / 11.9**
Portuguese	**9.9 / 13.2**	**16.4 / 19.3**	**8.5 / 11.7**	**9.1 / 12.7**
Spanish	**4.2 / 11.7**	**11.0 / 15.9**	**3.3 / 9.8**	**3.6 / 10.7**
Average	**7.0 / 13.8**	**16.2 / 21.8**	**6.2 / 12.2**	**6.8 / 13.1**

For these experiments, the resulting Word Accuracy is rather high, as can be noticed in Table 1, thus for many languages few points, corresponding to the errors, can be represented in the DET curves, that look stair-shaped. An example of the DET plots of the acoustic and confidence measures for the OOV experiments in Catalan language is shown in Figure 3 together with the DET curve resulting from the linear combination of the two confidences, with factors 0.75 and 0.25 for the acoustic and duration confidence respectively. Similar plots are obtained in the experiments with the other languages. Although the duration based CM produces a DET curve significantly worse than the one obtained by the acoustic based confidence, the duration information is complementary to the acoustic one, and valuable, as shown by the combination of the two measures, which improves the overall system confidence scoring mechanism.

The comparison of the performance of different systems is summarized in Tables 2 and 3 using the Equal Error Rate (EER) point on the DET curve, for which the value of τ gives equal error rates for the False Reject Rate and False Accept Rate errors. Table 2 shows the EER of the In-Vocabulary experiments. The results are given for the clean and noise conditions. The first column presents the results of the acoustic confidence measure. The performance of the confidence score based on the duration model *including the application words* is given in column two, and their linear combination, representing the best case, in column three. Since the number of errors for these tests is small, it is not surprising that in some cases the duration confidence is better than the acoustic one. Although the EER of the duration confidence is often much higher than the acoustic confidence EER, the combination of the two scores improves the system EER. The combination of the confidence measure based on the duration model *excluding the application words* and the acoustic confidence is detailed in the last column of Table 2. In this worst case, the EER of the duration confidence increases by 50% on average, but the overall confidence remains in the range or better than the original acoustic confidence for most languages.

Table 3 shows the same information of Table 2, but refers to the statistically more significant OOV experiments. The DET curves for these tests were obtained by taking the confidence scores of the application words correctly recognized as the true samples, and the confidence scores of a set of ~1K OOV phonetically balanced words as the false samples. The contribution of the duration confidence, even in the worst case, excluding the clean English tests, is positive.

5. Conclusions

A duration model for the computation of a duration confidence score has been proposed based on the statistics of the duration of phones in large contexts. This information provides complementary knowledge with respect to the acoustic one, which can used to improve the measure of the system confidence.

The duration based confidence measure can be used for connected word recognition. However, while insertion errors typically produce an incorrect duration profile, deleted words, usually recognised by the silence model, are not detected by our duration model.

6. References

[1] J. Pylkkonen, M.Kurimo, “Duration Modeling Techniques for Continuous Speech Recognition”, Proc. Interspeech 2004, pp. 385-388, 2004.

[2] M.W. Koo, C.H. Lee, B.H. Juang, “Speech Recognition and Utterance Verification Based on Generalized Confidence Score*”, IEEE Trans. Speech and Audio Proc.*, Vol. 9, n. 8, pp. 821–832, 2001.

[3] S. Goronzy, K. Marasek, A. Haag, and R. Kompe, ”Phone Duration Based Confidence Measures for Embedded Applications”. ICSLP-2000, Vol. 4, pp. 500-503, 2000.

[4] J. Pinto, R. Sitaram, “Confidence Measures in Speech Recognition based on Probability Distribution of Likelihoods”, Proc. Interspeech 2005, pp. 3001-3004, 2005.

[5] D. Albesano, R. Gemello, F. Mana, “Hybrid HMM-NN Modelling of Stationary-Transitional Units for Continuous Speech Recognition”, Proc. Int. Conf. On Neural Information Processing, pp. 1112–1115, 1997.

[6] G. Williams, S. Renals, “Confidence Measures from Local Posterior Probability Estimates”, Computer Speech and Language, Vol. 13, pp. 395–411, 1999.

[7] G. Bernardis, H. Bourlard, “Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems”, ICSLP 1998, pp. 775–778, 1998.

[8] L. Gillick, Y. Ito, J. Young, “A Probabilistic Approach to Confidence Estimation and Evaluation”, Proc. ICASSP 1997, pp. 879–882, 1997.

[9] D. Colibro, L. Fissore, C. Vair, E. Dalmasso, P. Laface, “A confidence measure invariant to language and grammar”, Proc. Interspeech 2005, pp.1001-1004, 2005.

[10] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, “Classification and Regression Trees”, Chapman&Hall (Wadsworth, Inc.): New York, 1984.

[11] Available at http://catalog.elra.info

[12] A. Martin, G. Doddington , T. Kamm, M. Ordowski, M. Przybocki , “The DET curve in Assessment of Detection Task Performance”, Eurospeech-97, pp. 1895-1898, 1997.