WEIGHTING PHONE CONFIDENCE MEASURES FOR AUTOMATIC SPEECH RECOGNITION

Gies Bouwman, Lou Boves, Johan Koolwaaij

A^{2}RT, Dept. Language & Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{G.Bouwman, L.Boves, J.Koolwaaij}@let.kun.nl

ABSTRACT

One of the most useful applications of Confidence Measures (CMs) in Automatic Speech Recognition systems is early detection of incorrect recognition hypotheses. A purely acoustic basis for such a CM is particularly important when tracking errors resulting from Out of Vocabulary speech, background noise or keyword substitution. A commonly taken approach is to compute scores on subword units of the hypothesized words and combine them in a word score. This paper investigates the assumption that some subword types contain stronger distinctive properties than others. Therefore, their scores ought to have a higher contribution in the eventual word scores. Experiments in a connected digit recognition task showed a relative Confidence Error Rate improvement of 6% on word level and 11% on sentence level in comparison to the baseline CM, with equal contribution of the phone confidence scores.

1. INTRODUCTION

For user friendly human-machine dialogs reliable estimates of the confidence with which a user utterance has been recognized are essential. This is as true for connected digit recognition (CDR) as it is for other tasks using speech input. Digit strings make for a relatively simple recognition task, because the vocabulary is rather small. On the other hand, in recognizing digit strings the linguistic constraints are minimal. Therefore, there is an obvious need for confidence measures which are almost completely based on acoustic information.

The ideal word-based acoustic confidence measure is without doubt the posterior acoustic probability:

$$P(W \mid X) \tag{1}$$

with W the hypothesized words and X the sequence of acoustic feature vectors. By definition, it is impossible to know a posterior probability in advance, so usually (1) is rewritten in the form of prior probabilities, by making use of Bayes' rule:

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)} \tag{2}$$

For recognition purposes, where it is the goal to find the words W that maximize *P(W|X)*, it is sufficient to find the W that maximize the numerator of the right hand term of (2). Besides the fact that the denominator is independent of W, it is nearly impossible to collect sufficient training data to make a fair estimation of P(X), so it is usually not considered. An important consequence, however, is that the resulting approximation can no longer be interpreted as an indication of the probability that the recognition result is correct. This is a problem when estimating a score for the acoustic confidence, which one would like to be an absolute and context independent value. The problem has been approached in several ways, like [1] in which a discriminatively trained hybrid HMM/ANN system was used to estimate the posterior probabilities directly. In this work we will use a Likelihood Ratio (LR) based method, which provides for an internal normalization of P(X):

$$LR(W \mid X) = \frac{P(W \mid X)}{P(\overline{W} \mid X)} = \frac{P(X \mid W)}{P(X \mid \overline{W})} \cdot \frac{P(W)}{P(\overline{W})} \cdot \frac{\cancel{P(X)}}{\cancel{P(X)}} \tag{3}$$

Where $\overline{W}$ is the hypothesis that any word(s) but W were realized. Since our current interest is in an acoustic confidence measure, we restrict ourselves to the left factor of the right hand term of (3) in the remainder of this paper.

In order to estimate the probability $P(X|\overline{W})$, it is required that a so-called anti-model is trained for every unit W. In the literature a number of approaches are described for selecting the speech material on which to train the anti-models (e.g. [4], [6]). Usually anti-models are trained on tokens that caused substitution errors. Quite naturally, these tokens are in some sense acoustically close to the words that should have been recognized.

In this paper we introduce a new way of selecting the speech for training the anti-model, that is based on a combination of a data driven and a rule based approach.

In large vocabulary Automatic Speech Recognition (LV-ASR), confidence scores for words must be derived from the contributions of the subword units. Although it would be possible to compute word level confidence scores for digits directly, we want to derive our scores from subword units, so as to allow the approach to generalize to LV-ASR. In a subword based approach, usually all units make an equal contribution to the eventual word confidence score (e.g. [1], [2]). This paper presents a study to validate the assumption that some subwords possess more discriminative ability than others and thus ought to have a higher contribution.

We apply the conventional two-pass scoring procedure: in the first stage a recognition is performed, on the basis of which each time frame of the input utterance is given a phone label. In the second pass, the labeled feature vectors are scored by their respective set of dedicated verification models to compute frame-based likelihood ratios. These are then propagated to a phone level normalization and eventually to word confidence scores.

Concerning this last step we will examine the following hypotheses in this work:

$H^{(1)}$ Normalizing phone confidence scores improves their discriminative capability, resulting in a reduction of the Confidence Error Rate (CER, see [7])

$H^{(2)}$ The CER can be reduced further by emphasizing the scores of the most discriminative phones of each word separately.

This paper is organized as follows. In Section 2 we discuss in general terms the training and scoring phase of the frame-based likelihood measure and propose a procedure to assess weighting coefficients for phone confidence scores. In Section 3 we describe the experiments to validate our hypotheses and present the results we obtained. Finally, in Section 4, we summarize and discuss our most important findings.

2. WEIGHTING PHONE CONFIDENCE

2.1 Training target- and anti-models

Recent studies have shown that phone likelihood ratios seem to perform best when the anti-model is trained on speech that gave rise to confusions. However, there seem to be two different approaches for selecting training tokens for the anti-model: (1) There are rule-based methods, in which all subword units are treated in the same way. For instance, [6] compared the performance of anti-models trained on material from instances of all other subword units, from instances of the same subword unit in other contexts, and from combinations of these selection rules. (2) There are also completely data driven approaches, that select only those instances where previous recognition has failed, e.g. [4]. When training anti-models for a small vocabulary task like CDR a third, intermediate, option is available: for our work we have derived selection rules from word level confusion matrices. This method combines the advantages of both approaches: the confusion matrices show real confusions; therefore we know the words which are relatively easy to confuse. The rules, however, generalize for the instances, making the procedure more robust for mismatch between train and test conditions. In this way we intend to maximize the discriminative power of the anti-models.

For the CDR experiments described in this paper, we used the confusion matrix of an experiment with an existing HMM phone based digit recognizer. The ten digits of Dutch can be described by means of 18 context independent phones. It appeared that most confusion occurred between the four digits (and one variant) with the vowel /e:/. Therefore, the transcription of the training corpus was adapted in such a way that digit dependent variants were created for the vowel /e:/. Although we used the original 18 models as our target-models, this allowed us to train 22 digit dependent anti-models for the most confusable phones.

The training procedure of anti-models is as follows. The recognizer is used to assign a phone label to each frame of the train material by forced alignment. Then, for each phone class, the labels are substituted by one of three classes: 'target', 'anti' and 'garbage', according to the selection rule of the phone under consideration. A single state anti-model is trained on all acoustic vectors with an 'anti' label. No intermediate alignments are to be made during this training procedure. In our approach, the target models are just the recognition models.

2.2 Frame-based phone scores

The target- and anti-models are used to score a corpus that has been recognized in a conventional way. First, the speech recognizer is used to obtain a phone level segmentation. Next, we compute the likelihood scores of the target and anti phone by means of their scoring models. Finally, a likelihood ratio score for each frame is computed.

Now that each frame has a phone dependent likelihood ratio, the scores must be combined into a word score. As in [1] we do this through an intermediate phone level normalization. A state level normalization might be an attractive alternative for its local specialization. However, stability of the scores becomes a crucial factor when the scores are propagated to higher levels, as a state score could have been based on a single frame. For a task like CDR, one might also choose to combine frame scores to word scores directly.

The raw phone-based score used in this work is the arithmetic average of the frame likelihood ratio scores:

$$CM(p_i) = \frac{1}{t_e - t_s + 1} \sum_{j=t_s}^{t_e} LR_i(j) \qquad (4)$$

where $CM(p_i)$ is the confidence score for phone p_i., to which frames t_s up to t_e were assigned. $LR_i(t_s)\ldots LR_i(t_e)$ are the likelihood ratio scores of these frames, as computed with the target and anti-model of p_i. (4) implements time normalization intrinsically. Sigmoid transformations of $CM(p_i)$ did not have much effect on the performance.

2.3 Discriminative ability

In this paper we want to investigate whether the confidence scores of some phones have more discriminative power than those of others; in other words, they are better predictors whether the word was realized or not. We will refer to this as the *discriminative ability* of the phone.

Assessing the discriminative ability of each phone is not trivial. In this study, we base it on the prior distribution of phone scores in correctly and incorrectly recognized words. To this aim we score a recognition result from some experiment on a development corpus, using the procedure as described in Section 2.2. For all phones $p_w(i)$ (the i-th phone of digit w, shorthand notation p_i), we split the confidence scores into two categories $I_{w,i}$ and $C_{w,i}$:

$$I_{w,i} = \{CM(p_i) \mid w^{inc}\} \quad C_{w,i} = \{CM(p_i) \mid w^{cor}\} \qquad (5a, 5b)$$

which are the sets of phone scores of incorrectly (w^{inc}) and correctly (w^{cor}) recognized digit tokens. Next, we can compute their means μ_C and μ_I and standard deviations σ_C and σ_I to express

$$Z_{w,i} = \frac{\left|\mu_{I_{iw,}} - \mu_{C_{w,i}}\right|}{\sigma_{C_{w,i}} + \sigma_{I_{w,i}}} \qquad (6)$$

$Z_{w,i}$ (shortly Z_i) represents the distance between the means of the two sets in terms of their standard deviations. The distribution of the scores is assumed to be normal, and therefore Z_i is a measure for the amount of overlap of I_i and C_i. If this overlap is small, we are able to separate the sets quite accurately by a

threshold that is set a priori. In other words, Z_i represents the discriminative ability of $CM(p_i)$.

2.4 Weighting coefficients

We investigate the assumption that phone confidence should not contribute in equal proportion to the final word confidence. It seems plausible to weigh each phone score with a coefficient that is related to the discriminative ability of the phone score. So we propose a word dependent vector of weighting coefficients, R_w for each word w

$$R_W = [r_1 ... r_M] \tag{7}$$

$$\sum_{i=1}^{M} r_i = 1 \tag{8}$$

where w consists of M phones. The word confidence score CM(w) then becomes:

$$CM(w) = \sum_{i=1}^{M} r_i \cdot CM(p_i) \tag{9}$$

The r_i in (9) must depend on the discriminative ability of the phones. Thus we define

$$r_i = \frac{Z_i{}^{\lambda}}{\sum_{i=1}^{M} Z_i{}^{\lambda}} \tag{10}$$

Exponent λ controls the relative contribution of the most discriminative phone score(s). If $\lambda = 0$, all phones have equal contributions. If $\lambda = 1$ the contributions of the phones are determined by their relative discriminative ability in the word. For $\lambda \rightarrow \infty$ the most discriminative phone dominates the confidence score for the words.

2.5 Accept or reject?

In order to make accept/reject decisions, we need to determine some threshold τ to compare the confidence score to. The decision then becomes

$$CM(w) \begin{matrix} > \\ \leq \end{matrix} \tau \Rightarrow \begin{matrix} accept & w \\ reject & w \end{matrix} \tag{11}$$

Setting a threshold introduces two kinds of errors. The first, False Accept Rate (FAR) is the rate of falsely accepted items. The second, False Reject Rate (FRR) is the rate of falsely rejected items. In this study, the threshold is optimized by minimizing the confidence error rate (CER_α):

$$\min_{\arg \tau} \; CER_\alpha(\tau) = \alpha \cdot FAR(\tau) + (1-\alpha) \cdot FRR(\tau) \tag{12}$$

in which α controls the *operating point* (FAR, FRR). If α is set to be the a-priori word recognition accuracy, then CER_α is actually the CER as proposed in [7]. In this way we find the threshold that is the best compromise between insertions and substitutions on the one hand, and deletions on the other.

2.6 Experiment setup

Experiments were carried out on a corpus consisting of three Dutch spoken connected digit databases: Polyphone, SESP and Casimir. All these corpora contain telephone speech recorded in a wide variety of acoustic conditions. The acoustic features were 14 Mel-scale Frequency Cepstrum Coefficients (c0 …c13), and their first and second order derivatives, i.e. 42 features. These vectors were based on 16 ms frames and a 10 ms frame shift. Phone HMMs were trained in a conventional way. Each state pdf was a mixture of maximally 32 Gaussian densities. The training set consisted of 9753 utterances with an average of 6.3 digits per utterance.

The weighting coefficients and threshold value were optimized on the recognition result of an independent development corpus (9155 digit strings), according to the proposed procedure. The test results were obtained on an unseen test corpus of 10.000 utterances (76682 digits).

The baseline performance of the models on the development set was 2.78% receiver Word Error Rate (WER_{rec}) (i.e. number of substitutions + insertions / total number of recognized words) and 15.3% receiver Sentence Error Rate (SER_{rec}) (i.e. the number of incorrectly recognized utterances / total number of utterances). These receiver error rates were used as the prior probabilities α in (12) for the evaluation, by computing $CER_{0.0278}$ and $CER_{0.153}$ at word and sentence level respectively.

3. RESULTS

This section reports on the experiments done with the proposed Confidence Measure for Utterance Verification, where it is the system's task to decide on whether to accept or reject the recognized digit.

λ	hard error rate (in %)			soft error rate (in %)		
	$CER_{0.0278}$	FAR	FRR	$CER_{0.0278}$	FAR	FRR
= 0	2.72	88.19	0.28	2.71	88.45	0.26
= 1	2.60	77.78	0.45	2.58	83.08	0.28
$\rightarrow\infty$	2.70	95.72	0.05	2.63	88.82	0.16
opt	2.55	78.10	0.38	2.53	83.19	0.22

Table 1. Error rates for λ=0, λ=1, $\lambda\rightarrow\infty$ and λ optimized per digit

Table 1 shows the False Accept Rate (FAR), False Reject Rate (FRR) and Confidence Error Rate (CER) for different values of λ on word level. The left hand part of the table contains the hard decision error rates, where all decision parameters, including threshold T of (11) and (12), are estimated on the independent development corpus. The numbers in the right hand part are the results of threshold optimization on the test set.

From Table 1, it is clear that weighting the phone scores according to their discriminative ability (λ=1) outperforms the baseline situation of equal contribution (λ=0) significantly. It yields a relative word level CER improvement of 4%. Therefore, the first hypothesis is strongly supported. The third row shows the opposite extreme, where the score of the most discriminative phone has an exclusive vote in the confidence score of the whole word. Obviously, the CER did not drop below the baseline situation. The row labeled 'opt' represents the performance of the confidence measure with a weighting vector with λ-values optimized per word, i.e. minimal word-dependent CERs. Each digit required a different optimal value,

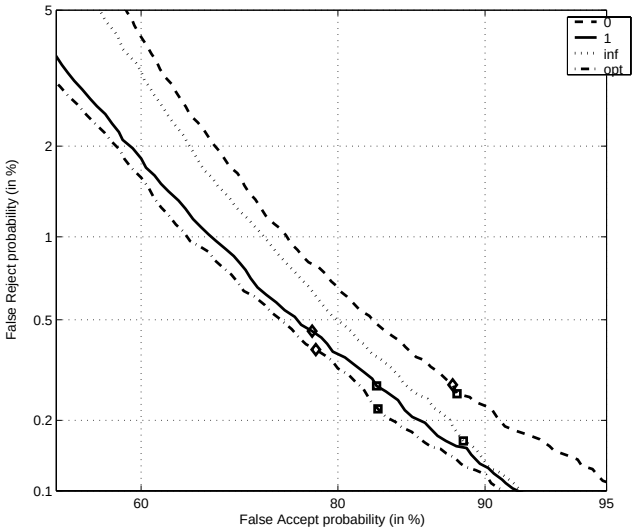

Figure 1. Detection Error Trade-off (DET)-curves for λ=0, λ=1, $\lambda \rightarrow \infty$ ("inf") and λ optimized per digit with hard decision (♦) and soft decision (■) operating points.

ranging from 0.75 to 6. This fine-tuning resulted in a further relative improvement of 2% relative.

Comparison of hard and soft decision error rates gives insight in the method's sensitivity to threshold optimization on a particular test set. Although the FAR and FRR have rather different values, it seems that the CER has been estimated quite accurately. There is, however, an exception for the case $\lambda \rightarrow \infty$, where the soft decision CER is almost 3% lower than the hard decision CER.

Finally, we report that the sentence level $CER_{0.153}$ decreased by 12%, viz. from 18.8% (for λ=0) to 16.5% (for λ optimized).

Figure 1 shows the Detection Error Trade-off (DET, see [3]) curve around the optimization points of Table 1. As can be seen, the improvement resulting from phone score weighting holds over a broad score domain around the operating points.

4. DISCUSSION

The evaluation of the experiments clearly shows that phone confidence scores make different contributions to the word confidence score. Discriminative ability has proven to be a strong criterion to base the weighting coefficients upon. It has also become apparent that emphasizing these coefficients, by fine-tuning λ of (7) per word gives an additional CER improvement.

It is remarkable that the CM that was solely based on the most discriminative phone score ($\lambda \rightarrow \infty$) seems to outperform the measure in which all phone scores were weighted equally (λ=0). After all, decisions based on more information are usually better. One possible explanation is that the phones always appear in the same digit context. The score may be about more than just the phone itself; the score implicitly takes some of the context into account as well. Still, as we already noted in the previous section, this score seems to be more sensitive to threshold optimization on the specific test set. Research in a large vocabulary task environment is expected to shed more light on this matter.

The hypotheses we stated are –at least in principle- independent of the task domain. Hence, it will become interesting to scale up the finding that the hypotheses are validated for a CDR task to the LV-ASR domain. This implies two major challenges:

1. generalization for our selection rules for the training material of the anti-models. A mixed approach of data-driven and rule-based criteria is founded on observed and therefore realistic errors on the one hand, yet is general enough to take account for avoiding potential errors on the other.
2. assessment of the discriminative ability of each phone confidence score. Further research is required to develop the proposed method from a word dependent level to the level of subword units that are less task-dependent.

5. CONCLUSION

Summarizing our most important results, we can conclude that

- we have proposed the idea for a new, semi-data driven approach of selecting training material for anti-models by making use of a word confusion matrix,
- we have verified the hypothesis that some phone scores possess more distinctive properties than others. A word confidence measure should rely heavier on the phone scores with a higher discriminative ability,
- we have presented a method to determine the discriminative ability of phone confidence scores in the small vocabulary ASR domain, and
- applying these findings in a Connected Digit Recognition task gave a relative Confidence Error Rate improvement of 6% at word level and 11% at sentence level.

6. REFERENCES

[1] Bernardis, G., Bourlard H., "Improving Posterior based Confidence Measures in HMM/ANN Speech Recognition Systems", *Proc. ICSLP '98*, Sydney, vol 3., pp. 775-778

[2] Garcia-Mateo C., Reichl W., Ortmanns S., "On Combining Confidence Measures in HMM-based Speech Recognizers", *Proc. ASRU '99*, Keystone, pp. 201-204

[3] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M., "The DET curve in Assessment of Detection Task Performance", *Proc. Eurospeech-97*, Rhodes, vol. 4, pp. 1895-1898

[4] Moreau N. and Jouvet D. "Use of a confidence measure based on frame level likelihood ratios for the rejection of incorrect data". *Proc. Eurospeech-99*, Budapest, vol. 1, pp. 291-294.

[5] Rahim M, Lee C.-H., Juang B.-H., "Discriminative Utterance Verification for Connected Digit Recognition". *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 266-277, May 1997

[6] Ramesh P., Lee C.-H., Juang B.-H., "Context Dependent Anti Subword Modeling for Utterance Verification", *Proc. ICSLP '98*, Sydney, vol. 7, pp. 3233-3236

[7] Wessel F., Macherey K., Schlüter R., "Using Word Probabilities as Confidence Measures", *Proc. ICASSP '98*, Seattle, vol. 1, pp. 225-228