POSTERIOR-BASED CONFIDENCE MEASURES FOR SPOKEN TERM DETECTION

Dong Wang[1], *Javier Tejedor*[1,2], *Joe Frankel*[1], *Simon King*[1] *and Jose Colás*[2]

1. The Centre for Speech Technology Research,
University of Edinburgh, UK
2. Human Computer Technology Laboratory,
Escuela Politecnica Superior, Universidad Autonoma de Madrid, Spain

ABSTRACT

Confidence measures play a key role in spoken term detection (STD) tasks. The confidence measure expresses the posterior probability of the search term appearing in the detection period, given the speech. Traditional approaches are based on the acoustic and language model scores for candidate detections found using automatic speech recognition, with Bayes' rule being used to compute the desired posterior probability.

In this paper, we present a novel direct posterior-based confidence measure which, instead of resorting to the Bayesian formula, calculates posterior probabilities from a multi-layer perceptron (MLP) directly. Compared with traditional Bayesian-based methods, the direct-posterior approach is conceptually and mathematically simpler. Moreover, the MLP-based model does not require assumptions to be made about the acoustic features such as their statistical distribution and the independence of static and dynamic co-efficients. Our experimental results in both English and Spanish demonstrate that the proposed direct posterior-based confidence improves STD performance.

Index Terms— Spoken term detection, confidence measure, posterior probabilities, MLP

1. INTRODUCTION

We adopt the definition of the spoken term detection (STD) task provided by NIST in 2006 [1], in which the goal is to locate spoken terms from audio archives precisely and reliably. Applications of STD include indexing and searching the increasing amounts of audio material available on the world-wide web, as well as in scenarios such as meetings, lectures, presentations and everyday conversation.

A typical STD system consists of three main components: a speech recogniser to transcribe input speech in terms of word or sub-word lattices; a lattice searcher to detect all potential occurrences of the search term; a decision maker to select only reliable detections.

Although more accurate speech recognition and better lattice searching schemes will generally improve the performance of any STD system, they are of little use without a good confidence measure. Many acoustic and linguistic cues could be used to calculate the confidence in each potential occurrence of the search term, e.g., whole-term or framewise acoustic likelihood, likelihood ratio of the top and second candidates, etc. Furthermore, various confidence measures can be combined to form a *composite* confidence measure.

Although various confidence might work well in practice, we prefer a posterior-based confidence as this is more theoretically clear. Basically, to tell what the spoken word is in the speech segment from t_1 to t_2, the decision based on the posterior probability $p(K_{t_1}^{t_2}|O_1^T)$ would be optimal in the sense of error-minimum, where $K_{t_1}^{t_2}$ denotes the event that term K appears in the speech segment from frame t_1 to t_2, and O_1^T is the whole speech with T frames. It is natural and straightforward to read $p(K_{t_1}^{t_2}|O_1^T)$ as confidence of the event $K_{t_1}^{t_2}$.

Conventionally, the posterior probability is computed from the acoustic and language model scores computed by the recogniser [2]. There are two shortcomings of this Bayesian-based approach: (1) the likelihood is computed from a generative probabilistic model, i.e., HMMs, which makes some incorrect assumptions, such as framewise and possibly component-wise independence of acoustic features, and a finite number of Gaussian mixtures; (2) computing the confidence is expensive and requires evidence from the whole lattice.

We propose a new posterior-based confidence measurement, which is directly calculated from posterior probabilities produced by a MLP network directly, so can be called a *direct posterior confidence measure*. The new approach does not make any assumptions regarding the distributional form and independence properties of the acoustic features and it requires no evidence from the lattice.

This new measure is inspired by the Tandem HMM-ANN hybrid architecture for speech recognition [3], but we use the posterior probabilities generated from the MLP as confidence measures instead of as observations for HMMs. MLP-based posteriors have also been used to re-score hypothesis in continuous speech recognition [4].

In the following section, we first review the conventional Bayesian confidence measure, and then in section 3 we present the direct posterior-based confidence in detail. The experiments will be presented and discussed in section 4, and conclusions drawn in section 5.

2. BAYESIAN-BASED POSTERIOR CONFIDENCE

For posterior-based confidence measurements, the posterior probability $p(K_{t_1}^{t_2}|O_1^T)$ is regarded as the confidence of search term K appearing in the speech signal from frame t_1 to t_2. According to the Bayesian formula, this posterior can be written as a product of conditional and prior probabilities:

$$p(K_{t_1}^{t_2}|O_1^T) = \sum_{\alpha,\beta} p(K_\alpha K_{t_1}^{t_2}, K_\beta|O_1^T) \quad (1)$$

$$= \sum_{\alpha,\beta} \frac{p(K_\alpha K_{t_1}^{t_2}, K_\beta, O_1^T)}{p(O_1^T)} \quad (2)$$

$$= \sum_{C_K} \frac{p(O_1^T|C_K, K_{t_1}^{t_2}) p(C_K, K_{t_1}^{t_2})}{p(O_1^T)} \quad (3)$$

where $K_{t_1}^{t_2}$ is an occurrence of the search term starting at frame t_1 and ending at frame t_2. K_α and K_β are any possible phone strings before and after $K_{t_1}^{t_2}$, with K_α starting at frame 1 and K_β ending at

frame T. To avoid cluttering, K_α and K_β are merged into C_K in Equation 3, representing the context of $K_{t_1}^{t_2}$.

In Equation 3, the conditional probability $p(O_1^T|K_{t_1}^{t_2}, C_K)$ is the acoustic likelihood, and the prior $p(K_{t_1}^{t_2}, C_K)$ is usually provided by the language model. The denominator $p(O_1^T)$ a constant. The Baum-Welch algorithm is usually employed to make computation of $p(O_1^T|K_{t_1}^{t_2}, C_K)$ efficient, so we denote this precise posterior-based confidence measurement the *Baum-Welch confidence*. A further reduction in computational cost can be achieved by replacing the the sum over all C_K with the single best path, as in equation 4:

$$p(K_{t_1}^{t_2}|O_1^T) \approx \frac{max_{C_K}\, p(O|K_{t_1}^{t_2}, C_K)p(K_{t_1}^{t_2}, C_K)}{max_{K_1^T} p(O|K_1^T)p(K_1^T)} \qquad (4)$$

This approximate confidence can be computed using the Viterbi algorithm and thus we denote it the *Viterbi confidence*. Although only an approximation, we observed no degradation in performance compared to the Baum-Welch confidence in our experiments. Figure 1 illustrates the computation of the Viterbi confidence for a potential detection of the search term *google*.

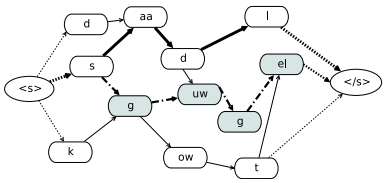

Fig. 1. Computing the Viterbi confidence for a candidate detection of the search term *google*. The thick solid lines indicate the globally-best path through the lattice and the thick dot-dash lines indicate the best path containing the pronunciation of *google*. A dot line represents all possible paths between its two ends. The Viterbi confidence is the ratio of the scores of these two paths.

3. DIRECT POSTERIOR-BASED CONFIDENCE

3.1. MLP-based posterior probabilities

It is well known that a standard 2-layer MLP network with softmax output activation can be used to estimate class posterior probabilities for a classification task. MLPs have been widely used in this fashion for speech recognition, by estimating the posterior probabilities for phone classes, given acoustic features as input [3]. Here, we use an MLP to estimate the posterior probability $p(Q_t|O)$ for each frame t, where Q_t is the phone class of the search term K at frame t. Q_t is obtained from the sub-word unit lattice produced by the recogniser. The structure of the MLP is shown in Figure 2.

3.2. Phone-independent posterior

Once we have the framewise posterior probability $p(Q_t|O)$, the search term confidence is calculated simply by summing the frame confidences, as shown in Equations 5-6. This confidence measure is independent of the context C_K, and only concerns acoustic properties. The MLP input is a window of $2W + 1$ frames of acoustic features. $W = 4$ in our experiments, meaning a 9-frame input window.

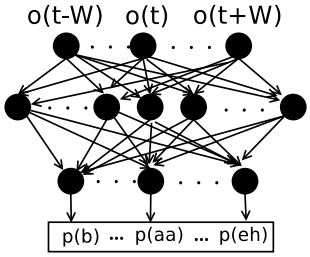

Fig. 2. The MLP network for framewise posterior probability estimation. The input layer consists of 9 frames, amounting to 351 input nodes, and the outputs are phone categories, which for English include 40 vowels and consonants plus a short and a long silence. The hidden layer, whose size is optimised by cross-validation, contains 5k hidden units. For Spanish, the output layer has 47 phones plus a short silence and a long initial and a long final silence, and the hidden layer, whose size is also optimised to cross-validation, contains 1k hidden units.

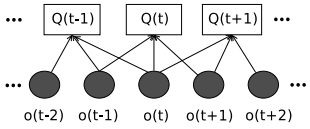

Fig. 3. The graphical representation of the phones-are-independent model for posterior confidence calculation. $Q(t)$ is the phone class at frame t, and $o(t)$ is the observed acoustic feature at time t.

$$p(K_{t_1}^{t_2}|O_1^T) = \prod_{t=t_1}^{t_2} p(Q_t|O_1^T) \qquad (5)$$

$$= \prod_{t=t_1}^{t_2} p(Q_t|o_{t-W}, ..., o_t, ...o_{t+W}) \qquad (6)$$

3.3. Phone-dependent posterior

The strong phones-are-independent assumption above means that some useful information from linguistic constraints is not used. To remedy this, dependence should be added between phones, as shown in Figure 4. We tried two ways of implementing the phone dependency: *direct LM score integration* and *Baum-Welch LM posterior*.

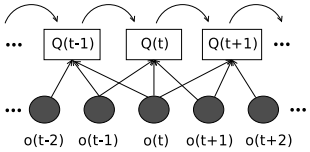

Fig. 4. The graphic representation of the phones-are-dependent model for posterior confidence calculation. $Q(t)$ is the phone at frame t, and $o(t)$ is the observed acoustic feature at time t. In this model, the phone dependency is described by a bigram LM.

3.3.1. Direct LM score integration

To model linguistic constraints, we introduce the variable K^l which represents the search term K in the word layer. If we assume that K^l determines K's phonetic form, i.e.,$p(K^l, K) = p(K^l)$ and that,

given phonetic form K, K^l is independent of acoustic observation O, i.e., $p(K^l|K,O) = p(K^l|K)$ then there is a *best* context C'_{K^l}, which accounts for most of the probability mass of the accumulated linguistic evidence, i.e., $\sum_{C_{K^l}} p(K^l|C_{K^l})p(C_{K^l}) \approx p(K^l|C'_{K^l})$ then the posterior probability of a detection will be the product of the acoustic score and LM score, as shown by Equation 7-10.

$$p(K,K^l|O) = \sum_{C_{K^l}} p(K,K^l,C^l_K|O) \quad (7)$$

$$= p(K|O)\sum_{C_{K^l}} p(C_{K^l},K^l|O,K) \quad (8)$$

$$= p(K|O)\frac{\sum_{C_{K^l}} p(K|C_{K^l})p(C_{K^l})}{p(K)} \quad (9)$$

$$\approx p(K|O)\frac{p(K^l|C'_{K^l})}{p(K)} \quad (10)$$

Note that the LM score $p(K^l|C'_{K^l})$ has been stored in the lattice and the unigram probability $p(K)$ can be obtained by table look-up, therefore the searching procedure is the same as for the phone-independent case.

3.3.2. Baum-Welch LM posterior

To integrate linguistic constraints into the term confidence, we can also regard the confidence estimation as a two-step process: in the first step, only *acoustic* confidence is considered, which comes from the MLP structure; in the second step, we test the *linguistic* confidence assuming all allowable phone alternatives have been included in the phone lattice. Finally the real confidence is computed as the product of the acoustic and linguistic posterior. This approach can be formulated as Equation 11-13, where L denotes the entire phone lattice.

$$p(K,K^l|O) = p(K|O)p(K^l|L) \quad (11)$$

$$= p(K|O)\frac{p(K^l,L)}{p(L)} \quad (12)$$

$$= p(K|O)\frac{\sum_{C_{K^l}} p(K^l,C_{K^l})}{p(L)} \quad (13)$$

Note that the linguistic confidence $p(K^l|L)$ is a true probability, and relates to linguistic constraints only. In addition, this is a *global* score and therefore requires a forward-backward computation. For this reason, we call $p(K^l|L)$ a Baum-Welch LM confidence.

4. EXPERIMENTS

We tested the proposed confidence measurement on both English and Spanish tasks: in English, the test was performed on data from the meetings domain; for Spanish, the test data is read speech.

In all experiments, we used the HTK toolkit to build the acoustic models. The *Lattice2Multigram* tool from Brno University of Technology (BUT) was used for term searching and Bayesian-based confidence measuring. The ICSI toolkit QuickNet was used for MLP training and posterior generation.

Standard 13-dim MFCCs plus their first and second derivatives (39-dim vectors in total) were used as acoustic features for the HMMs and standard 13-dim PLPs plus their first and second derivatives were used as MLP input features.

Results will be presented in terms of average term weighted value (ATWV) as defined by NIST [1]. We also present detection curves (DET) to examine different operating points with various recall and false-alarm rates.

	Viterbi	Baum-Welch	Post(DLM)	Post(BW)
ATWV	0.57	0.57	0.55	0.59

Table 1. STD results in terms of ATWV with four confidence measurements. All these measuring are performed on the same lattices generated by a speech recogniser using 7-gram phone LMs.

4.1. English experiments

The first experiments are on meeting domain data for English. The acoustic model, based on triphone HMMs, was trained on over 100 hours meeting data which were collected from several sites. The independent headset microphone (IHM) channels were used. 7-gram phoneme LMs, which were trained on a text corpus of 51M words using the SRILM toolkit, were used for speech recognition and lattice generation. The results are reported on test data from the NIST Spring 2004 Rich Transcription (RT04s) evaluation, and the corresponding RT04s development set was used for parameter tuning.

For STD evaluation, we selected 90 words as search terms, including frequently used terms, people and city names, and some compound words. 45 of these words are out-of-vocabulary (OOV) words whose pronunciations were predicted by the letter-to-sound module of the Festival system [5]. More details of the experimental setup can be found in [6].

In Table 1, we present the experimental results in terms of ATWV with four confidence measurements:

- Viterbi: the most commonly-used Viterbi implementation of the Bayesian confidence measure, as per Equation 4.
- Baum-Welch: a standard Baum-Welch implementation of the Bayesian confidence measure, as per Equation 3.
- Post(DLM): direct posterior-based confidence measure with direct LM score, as per Equation 10.
- Post(BW): direct posterior-based confidence with Baum-Welch LM posterior, as per Equation 13.

The results presented in Table 1 show that the direct posterior-based confidence with Baum-Welch LM posterior achieved the best performance, with the two Bayesian approaches performing almost the same. The direct posterior confidence with direct LM scores did not work well, which we suppose is because a single LM score is not sufficient to account for the context information required in this experiment.

Figure 5 shows the DET curves of these four measures. Again, the direct posterior-based confidence with Baum-Welch LM posterior gave the best performance, and the two Bayesian-based confidence measures performed nearly the same.

4.2. Spanish experiments

For the Spanish experiments, we used the ALBAYZIN geographical domain database [7]. More details of this corpus can be found in [8]. For the STD evaluation, we selected 80 words in the geographical domain as our search terms, based on their high frequency of occurrence in the development and test sets. All of them are OOV words. In Table 2 we present the results for the same four confidence measures as for English. The results demonstrate that the direct posterior confidence with direct LM score gave the best performance.

Figure 6 shows DET curves for these four measures. It is seen that the direct posterior-based confidence with direct LM posterior gave the best performance. As in English, the two Bayesian-based confidence performed nearly the same.

	Viterbi	Baum-Welch	Post(DLM)	Post(BW)
ATWV	0.18	0.18	0.26	0.15

Table 2. STD results in terms of ATWV for four confidence measures. All results are obtained from the same lattices, which were generated by a speech recogniser using a 2-gram phone LM.

4.3. Discussion

By inspecting the DET curves in Figures 5 and 6, we observe that, although the direct posterior-based confidence performs better than the Bayesian-based confidence, it behaves differently in English and Spanish. In English, the Baum-Welch LM posterior helped improve the quality of the confidence. However, in Spanish, direct LM score integration performed better, and the LM posterior in fact reduced the performance. This discrepancy might stem from the differing length of linguistic context used: for English we used a 7-gram phoneme LM, while for Spanish we used a 2-gram. A short-span context may be well represented by a single LM score, while for a long-span context a real posterior considering competing paths would be helpful.

Another observation is that the benefits achieved by using the direct posterior approach (either with direct LM score or Baum-Welch LM posterior) are greater for Spanish than for English. This may be due to the fact that the experimental condition in Spanish is read speech while in English it is spontaneous conversation from meetings.

Finally, we mention that, although the Baum-Welch LM posteriors help improve the performance in the long linguistic context scenario, this requires a forward-backward computation. On the other hand, direct LM scores are easily integrated into the acoustic confidence, and no extra lattice computation is required. In the case of low-order LMs, this direct posterior confidence with direct LM scores is good enough, and more efficient.

5. CONCLUSIONS

We attribute the success of the MLP-based confidence to three things. (1) Acoustic confidence is a *local* score, which is highly dependent on the current frame and its near neighbours; the neighbouring frames are themselves highly correlated. The MLP network is able to model this frame-wise dependency. (2) With sufficient training data, the MLP structure can represent any posterior distribution, whereas Gaussian-based systems converge to the assumed model, instead of the true distribution. (3) Combined with the Baum-Welch LM posterior, the *local* acoustic posterior might be efficiently refined by the long-span linguistic confidence, thus becoming a *global* posterior.

6. ACKNOWLEDGEMENTS

DW is a Fellow on the EdSST interdisciplinary Marie Curie training programme. JT is a visiting researcher at CSTR. JF was funded by the Edinburgh Stanford Link. SK is an EPSRC Advanced Research Fellow. Igor Szoke and colleagues in the Speech Processing Group of FIT, Brno University of Technology provided the lattice search tools. This work used the Edinburgh Compute and Data Facility which is partially supported by eDIKT. Part of this work was funded by the Spanish Ministry of Science and Education (TIN 2005-06885).

7. REFERENCES

[1] NIST, *The spoken term detection (STD) 2006 evaluation plan*, National Institute of Standards and Technology, Gaithersburg, MD, USA, v10 edition, September 2006.

[2] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 633–636.

[3] H. Hermansky, D. P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP 2000*, Istanbul, Jun. 2000.

[4] G. Zavaliagkos, Y. Zhao, R. Schwartz, and J. Makhoul, "A hybrid segmental neural net/hidden markov model system for continuous speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, no. 1, pp. 151–160, January 1994.

[5] R. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, April 2007.

[6] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection.," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP08)*, March 2008.

[7] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mari no, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proc. Eurospeech*, September 1993, vol. 1, pp. 653–656.

[8] J. Tejedor, D. Wang, J. Frankel, S. King, and J.Colás, "A comparison of grapheme and phoneme-based units for spanish spoken term detection," *Speech Communication, In press.*, March 2008.

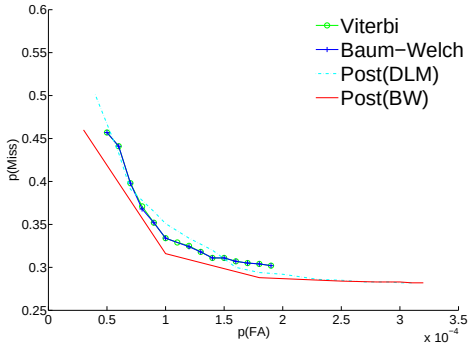

Fig. 5. DET curves for four posterior-based confidence measures. P(Miss) and P(FA) are the miss rate and false alarm rate respectively, as defined in the NIST STD 2006 evaluation plan [1].

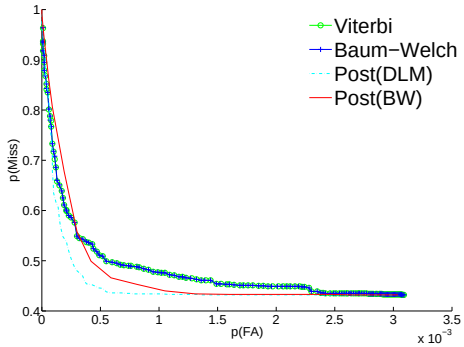

Fig. 6. DET curves for four posterior-based confidence measurements. The coordinates are the same as in the English experiment described earlier.