PHONE LEVEL CONFIDENCE MEASURE USING ARTICULATORY FEATURES

Ka-Yee Leung, Manhung Siu

Department of Electrical and Electronic Engineering,
Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong.
eekaren@ust.hk, eemsiu@ee.ust.hk

ABSTRACT

Confidence measures are used in a number of applications to verify the user input or to measure the certainty of the recognition outputs. Most of the HMM-based systems use MFCC features with Gaussian mixtures models to estimate confidence. In this paper, we propose a new approach to estimate confidence by combining the posterior probabilities of articulatory features (AF) computed by a set of AF classifiers. This AF-based confidence measure gives comparable performance in terms of classification equal error rate (EER) to the Gaussian mixture-based approach but reduces the computation by 50% (as measured by the approximated number of multiplications) and consumes smaller memory. When the AF-based confidence is combined with confidence from the Gaussian mixtures, the EER is further reduced. This AF confidence can be particularly useful for platforms with limited computing resources such as hand-held devices.

1. INTRODUCTION

Although current speech recognition systems can achieve high recognition accuracy, they are not yet perfect and errors do occur. In applications such as the automatic call attendant, it is better to identify a recognition error based on a measure of recognition certainty or **confidence** and in the case of low certainty, it can prompt the user to repeat the input rather than using the likely erroneous recognition result for the transaction. In other applications, such as computer assisted reading or pronunciation learning, confidence measures can be applied to verify the user inputs.

Because mel-frequency cepstral coefficients (MFCC) are commonly used for recognition, confidences are also commonly derived from MFCC features directly [1, 2, 3] using Gaussian mixture models (GMM) within the HMM framework. We called this the GMM-based confidence. In this paper, we propose an alternative approach that estimates confidence using articulatory features (AF), which captures the characteristics of speech production. These AFs can be viewed as an intermediate representation of the MFCCs because they are extracted from MFCCs by a set of classifiers as reported in our earlier work [4].

Articulatory information, such as the place and manner of articulations, is an abstract description of some important properties during speech production. The use of articulatory information as features is more phonologically meaningful than acoustic features because they explicitly represent the underlying speech production process. AF recognition systems, such as [5, 6], have been shown to work as good as the acoustic feature systems in terms of recognition performances. Using the extracted AF, confidences are computed by considering the correctness of various articulatory properties of the hypothesized phone which are directly related to the quality of pronunciation rather than solely on the acoustic properties. For applications such as pronunciation learning, not only can it indicates the confidence but also potentially suggests possible production errors. In addition, less computational cost and memory are consumed when compared to the confidence measure approaches using mixture models within HMMs.

In this paper, confidences are derived from either the acoustic feature models or the articulatory feature models without the use of any language modeling information. Instead of evaluating word confidences, we evaluate the phone-level confidences which can be used as important building blocks for word/utterance verification [7]. In the next section, we describe how to estimate posterior probability as confidence based on the MFCC features with mixture models and the articulatory features with classifiers. In Section 3, we describe the experiments and results. We conclude the paper in Section 4.

2. ESTIMATING CONFIDENCE

Suppose a recognizer outputs a string of hypothesized units during decoding (this string can be of either phones or words), it can be converted into a sequence of phones, denoted as $\{w_1, w_2, ..., w_M\}$, with the corresponding starting and ending times, $\{[ts_1, te_1]$, ..., $[ts_M, te_M]\}$. Phone level confidences are computed on this hypothesized phone string with the corresponding phone boundaries. To fairly compare different confidence measure approaches, a single hypothesized phone sequence and the corresponding phone boundaries are used. Given the phone labels and boundaries, we can actually consider the confidence of each phone independently. To simplify our notation, for phone w_i, its starting time is denoted as ts instead of ts_i and the ending time is denoted as te instead of te_i.

2.1. Estimating Posterior Probabilities

The posterior probability of a phone w_i starting at time ts and ending at time te can be written as $p(w_i|x_{ts}^{te})$, where x_{ts}^{te} is the acoustic observations (MFCC) from time ts to te. By applying Bayes' rule,

$$\log p(w_i|x_{ts}^{te}) = \log \frac{p(x_{ts}^{te}|w_i)p(w_i)}{p(x_{ts}^{te})} \quad (1)$$

where $p(x_{ts}^{te})$ is the total likelihood of the observations and $p(x_{ts}^{te}|w_i)$ and $p(w_i)$ are the observation likelihood conditioned on phone w_i

and phone prior probability respectively. Equation 1 is sometimes called the log likelihood ratio (LLR) because it is a ratio of the target phone likelihood against the observation likelihood and is commonly used as confidence [1, 2].

2.2. GMM-based Confidence

In an HMM-based recognition system, the likelihood of the target phone in the numerator of Equation 1, $p(x_{ts}^{te}|w_i)$, is often approximated by the likelihood of the best state sequence through the target phone. The denominator, however, represents the likelihood of the observations and has to be computed by explicitly summing over all possible state sequence of a filler model, such as an unconstrained phone loop or a loop of HMM phone-states.

$$p(x_{ts}^{te}) = \sum_{Q} p(x_{ts}^{te}|Q)p(Q). \quad (2)$$

This is again often approximated by the likelihood of the best path via the Viterbi algorithm, given by

$$p(x_{ts}^{te}) \approx \max_{Q} p(x_{ts}^{te}|Q)p(Q). \quad (3)$$

In this paper, we used a loop of HMM phone-states with the Viterbi approximation for both phone-conditional likelihood and the observation likelihood when computing the GMM-based confidence.

2.2.1. Computation

It is complicated to measure the exact amount of computation because it depends on the implementation and platform. To simplify our computation, we ignore the computation involved in performing Viterbi maximization and the cost of additions but instead, focuses on the approximated number of multiplications. Certainly, the use of pruning or tying and other techniques may also affect the results. Because they really vary depending on particular task and performance level, they are also ignored in the computation.

Suppose the GMM-based confidence is estimated using an unconstrained loop of L phone with K states per phone and each state uses M mixtures for a D-dimensional feature vector. To compute the likelihood of a D-dimensional diagonal Gaussian, it involves at least $2D$ multiplications because of the quadratic terms and the inverse covariance. The total number of multiplications needed to evaluate the likelihood of one frame is then approximately $2D \times M \times K \times L$. The memory usage is a function of the size of the state space of the order $O(L \times K)$

2.3. AF-based Confidence

2.3.1. AF Extraction

In this work, five abstract articulatory properties *voicing, rounding, front-back, manner and place* are used and their values are shown in Table 1.

Similar to our previous work [4], each phone is mapped to a deterministic vector of AF property values. This is just a rough approximation because of human articulators are actually asynchrony in real speech, however, the true AF properties are difficult to obtain. Let's denote $AF_k(w_i)$ as the k-th property of phone w_i. If we use the phone /aa/ as an example, $\{AF_1$(aa) = voiced, AF_2(aa) = not rounded, AF_3(aa) = back, AF_4(aa) = vowel, AF_5(aa) = low$\}$. That is, a phone is represented by a 5-dimensional target vector. Under this framework, the probability

Feature group	*Values*	*# Value*
Voicing	Voiced, Unvoiced, Sil	3
Rounding	Rounded, Not Rounded Nil, Sil	4
Front-back	Front, Back ,Nil, Sil	4
Manner	Vowel, Stop, Fricative, Nasal Approximant & Lateral, Sil	6
Place	High, Middle, Low, Dental, Labial, Coronal, Palatal, Velar, Glottal, Sil	10

Table 1. 5 articulatory properties and their feature values

of observing a phone is equivalent to the probability of observing its articulatory properties, that is,

$$\begin{aligned} p(w_i) &= p(AF_1(w_i), AF_2(w_i), AF_3(w_i), AF_4(w_i), AF_5(w_i)), \\ &\approx \prod_k p(AF_k(w_i)). \end{aligned} \quad (4)$$

by assuming the AF properties are independent to simplify the calculation. To be exact, not all the AF properties are independent.

The AF extraction process is the same as in [4]. For each articulatory property, say the k-th one, a Multi-Layer Perceptron (MLP) is trained to determine the posterior probabilities, $p_k(c|\hat{x}_t)$, of observing value c using the t-th window of MFCCs, $\hat{x}_t$ (a window of 9-frames MFCC around time t is used in AF models). For example, if k = 1 and c = voicing, $p_k(c|\hat{x}_t) = p_1$(voicing$|\hat{x}_t)$ is the probability that the speech is voiced at time t.

2.3.2. Confidence Using the Articulatory Features

Similar to what is defined in Section 2.2, our goal is to compute the posterior probability, $p(w = i|\hat{x}_{ts}^{te})$, in which w spans from ts to te. To clarify our discussion, we further denote w_t as the phone label for the t-th frame. That is, $w = i$ implies $w_{ts} = i, w_{t_s+1} = i, \ldots, w_{te} = i$. As shown in Equation 4, the probability of a phone in the AF framework is equal to the joint probability of its five AF properties. If these AF properties are assumed to be independent, the posterior probability $p(w_t = i|\hat{x}_t)$ can be approximated by,

$$p(w_t = i|\hat{x}_t) \approx \prod_{c=1}^{5} p(AF_c(w_t = i)|\hat{x}_t), \quad (5)$$

where $p(AF_c(w_t = i)|\hat{x}_t)$ is the output from the c^{th} AF MLP.

Using the AF probabilities, per-frame phone posterior probability, $p(w_t = i|\hat{x}_t)$, can be computed according to Equation 5. The next question is how to compute the posterior probability for a sequence of frames, $p(w = i|\hat{x}_{ts}^{te})$. While it is tempting to just multiply the per-frame phone posterior probability, the frames are not independent. A reasonable assumption would be that the observations are conditional independent given that we know the phone labels. In such case, the per-frame phone likelihoods, $p(\hat{x}_t|w_t = i)$ can be combined. This gives,

$$p(\hat{x}_{ts}^{te}|w_{ts}^{te} = i) = \prod_{t=ts}^{te} p(\hat{x}_t|w_t = i) \quad (6)$$

where the per-frame phone likelihood is obtained by applying Bayes rule

$$p(\hat{x}_t|w_t = i) = \frac{p(w_t = i|\hat{x}_t)p(\hat{x}_t)}{p(w_t = i)}. \quad (7)$$

To simplify the derivation, the phone prior probabilities are assumed to be uniform. Using Bayes rule, the posterior probability for phone i occurring between ts and te is given by,

$$\log p(w_{ts}^{te} = i|\hat{x}_{ts}^{te}) = \log \frac{p(\hat{x}_{ts}^{te}|w_{ts}^{te} = i)}{\sum_k p(\hat{x}_{ts}^{te}|w_{ts}^{te} = k)} \quad (8)$$

$$= \log \frac{\prod_t p(\hat{x}_t|w_t = i)}{\sum_k \prod_t p(\hat{x}_t|w_t = k)} \quad (9)$$

$$= \log \frac{\prod_t p(w_t = i|\hat{x}_t)p(\hat{x}_t)}{\sum_k \prod_t p(w_t = k|\hat{x}_t)p(\hat{x}_t)} \quad (10)$$

This can be further simplified by removing the terms $\prod_t p(\hat{x}_t)$ which appears in both the numerator and denominator. Applying the AF probabilities as given Equation 5, Equation 10 is expanded to

$$\log p(w_{ts}^{te} = i|\hat{x}_{ts}^{te}) = \log \frac{\prod_t \prod_c p(AF_c(w_i)|\hat{x}_t)}{\sum_k \prod_t \prod_c p(AF_c(w_k)|\hat{x}_t)}. \quad (11)$$

2.3.3. Computation

If articulatory features are already extracted, the multiplication required to evaluate AF-confidence is quite small. For a set of L phones, the number of multiplications per frame is only $4L$. However, the number of multiplications required to extract the AFs using the MLP classifiers is more substantial. Suppose each of the five AF MLPs has N inputs, R hidden units and G output units, the number of multiplications for each AF classifier is $N \times R + R \times G$. Again, we ignore the cost of additions in our computation. The memory usage for the AF is very small and it is only of the order of the number of hidden units, O(R).

3. EXPERIMENTS

All our experiments were performed on the TIMIT [8] corpus. 39-dimensional acoustic feature vector, 12 MFCC, the normalized power as well as their first and second order derivatives were used for recognition. Each of the 42 context independent phones was modeled by a 3-state left-to-right Hidden Markov Model (HMM) with 16 diagonal covariance Gaussian mixtures per state. For simplicity, this is called the 16 Gaussian mixtures models (GMM). The models were trained and tested on the SI and SX sentences of the TIMIT training and testing set respectively using the HTK [9].

For the AF extraction, nine frames of 26-dimension MFCC (no 2nd order derivatives) were used as input to the five AF MLPs. These MLPs contained a single hidden layer which was composed of 50 hidden units. They were trained using the Quicknet software [10].

To fairly compare different approaches, confidences were estimated on the hypothesized phones obtained from the 16 GMM. The hypothesized phones are first compared with the true transcription using dynamic programming alignment in order to mark the correctness of the hypothesized phone sequences. There are a total of 43973 hypothesized phones, 30164 are correctly recognized while 13809, including insertions, are incorrectly recognized.

In addition to the 16 GMM used in recognition, another context independent HMMs of the same topology with only 8 diagonal covariance mixtures per state were also used to investigate the effect of model resolution against the computational cost on the confidence performance. This is called the 8 GMM.

Confidence measures	*# Multiplication*	*EER %*
8 GMM	79000	36.9
16 GMM	158000	36.4
Using AF prob.	60000	36.3

Table 2. Number of multiplications and EERs of the three approaches

3.1. Evaluation Metrics

To evaluate the performance of the different confidence measure approaches, we use the classification equal error rate (EER) and the Detection Error Trade-off Curve (DET). The confidence of a phone is compared to a threshold. A phone is accepted if its confidence is larger than the threshold, otherwise, it is rejected. A false acceptance (FA) occurs if a mis-recognized phone is accepted and a false rejection (FR) occurs if a correctly recognized phone is rejected. At each threshold, the FA rate and FR rate are calculated in which FA rate equals to the number FA normalized by the total number of mis-recognized phones and FR rate equals the number of FR normalized by the total number of correctly recognized phones. The threshold used for classification is varied to obtain multiple sets of false acceptance (FA) rates and false rejection (FR) rates. These FR and FA rates are plotted in the DET [11] curve.

Besides the DET curve, the EER [3], i.e. the point on the DET curve in which the FR rate and the FA rate are the same, is also used for evaluation.

3.2. Evaluation Results

There are a total of three confidence measure approaches being evaluated, both the computational cost and the EER are compared among them. All the confidence measure approaches are based on the same hypothesized phone string and the corresponding durations obtained in HMM-based recognition.

These three approaches compute LLR from:

- acoustic feature-based 8 GMM
- acoustic feature-based 16 GMM
- the AF probabilities estimated by the five AF MLPs

We compare the computational cost of different approaches based on the number of multiplications required as described in Sections 2.2.1 and 2.3.3. As discussed in Section 2.2.1, only the number of multiplications is counted. The number of multiplications required to compute the posterior probability for each frame using the GMM was 42 phones $\times 3$ states per phone $\times$ # mixtures per state $\times 39$ MFCC $\times 2$. To estimate AF-based confidence, the multiplications required for each frame was 5 MLPs $\times 50$ hidden unit $\times$ (9 frame $\times 26$ MFCC + total 27 MLP outputs) + 42 phones $\times 4$. Table 2 summarizes the number of multiplications required per frame as well as the EERs of the above three confidence measure approaches on 1344 testing utterances with a total of 43973 hypothesized phones.

As shown in Table 2, using the models with 16 GMM to estimate confidences double the number of multiplications as compared to that of using 8 GMM. The number of multiplications required by the AF-based approach was approximately 24% and 62% less than those required by using 8 GMM and 16 GMM respectively and it is the lowest EER out of the three approaches.

Confidence measure performance can be improved by combining the GMM-based confidence with the AF-based confidence.

	Linear combination			
	Min	*Max*	*Sum*	*Mul*
EER	0.37067	0.3635	0.3573	0.3780

Table 3. EER of different linearly combined confidences from the 16 GMM and the AF-based approach

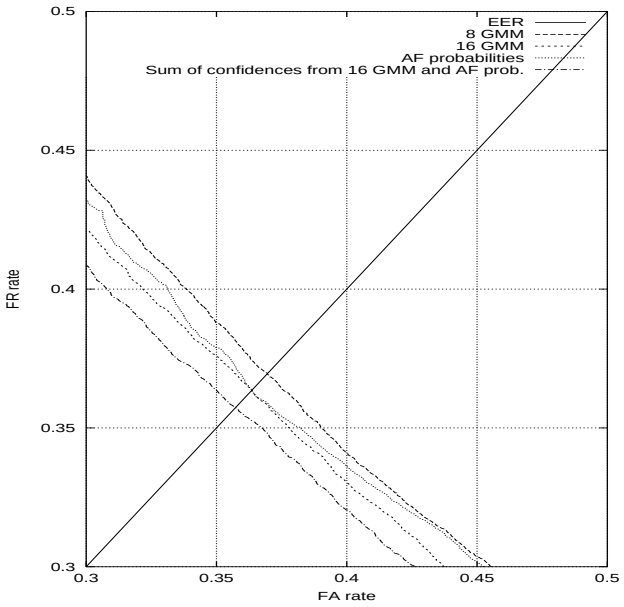

Fig. 1. DET plot of various confidence scoring approaches

Four different combination operators are applied to the 16-GMM-based confidence and the AF-based confidence. They are minimum (*Min*), maximum (*Max*), summation (*Sum*) and multiplication (*Mul*) operators. For the *Sum* and *Mul* combinations, different combination weights can be applied to the two confidences. Our previous experience in combining AF and MFCC systems [4] indicated that the weights are depended on the performances of the individual systems. Because the EERs of the two measures before combination are very similar, summation and multiplication with equal weights are applied. The EER of different combinations are summarized in Table 3 and a lower EER is obtained only for the *Sum* combination which is consistent with the findings in [5].

Figure 1 summaries the DET plots of the four approaches. As shown in the plot, confidences computed using the 8 GMM give the worse performance and consistently better results are obtained by the 16 GMM. Although nearly the same EER are obtained by the 16 GMM and the AF-based confidence, the 16 GMM performs slightly better at some other operating points. Summing the 16-GMM-based confidences and the AF-based confidences resulted in the best performance over all operating points.

4. CONCLUSION

In this paper, we have proposed a phone confidence measure approach based on the articulatory features probabilities estimated from MLP classifiers. Articulatory features have the advantage that it is phonologically meaningful. For applications such as pronunciation learning, not only can it indicates the confidence but also potentially suggests possible production errors. This approach gave similar performance compared to the commonly used approach of mixture models. In addition, by combining the mixture model and the AF-based confidences, improved performance over either one is obtained. The AF-based confidence requires significantly less computation and memory. The reduced computation and memory can be particularly useful for applications with limited computing resources, such as those in hand-held devices.

5. ACKNOWLEDGEMENTS

We would like to thank Katrin Kirchhoff for her valuable suggestions on the AF model and the Realization Group at ICSI, Berkeley USA, for providing us with the QuickNet software.

6. REFERENCES

[1] S. O. Kamppari and T. J. Jazen. Word and Phone level Acoustic Confidence Scoring. *Proceedings of ICASSP*, Volume III, pages 1799-1802, 2000.

[2] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig and A. Stolcke. Neural-Network Based Measures of Confidence for Word Recognition. *Proceedings of ICASSP*, Volume II, pages 887-90, 1997.

[3] F. Wessal, R. Schluer, K. Macherey and H. Ney. Confidence Measures for Large Vocabulary Continous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, Volume 9(3), pages 288-98, 2001.

[4] K. Y. Leung and M. Siu. Speech Recognition Using Combined Acoustic and Articulatory Information with Retraining of Acoustic Model Parameters. *Proceedings of ICSLP*, pages 2117-20, 2002.

[5] K. Kirchhoff, G. A. Gink and G. Sagerer. Conversational Speech Recognition Using Acoustic And Articulatory Input. In *Proceedings of ICASSP*, Volume III, pages 1435-1438, 2000.

[6] K. Erler and L. Deng. HMM Representation Of Quantized Articulatory Features For Recognition Of Highly Confusible Words. In *Proceedings of ICASSP*, Volume I, pages 545-548, 1992.

[7] R. A. Sukkar and C. H. Lee. Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. In *IEEE Transactions on Speech and Audio Processing*, Volume 4(6), pages 420-9, 1996.

[8] V. Zue, S. Seneff and J. Class. Speech Database Development at MIT: TIMIT and Beyond. *Speech Communication*, 9(4): 351-356, August 1990.

[9] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland.*The HTK Book for HTK 3.0*. Microsoft Corporation, July 2000.

[10] P. Frber. Quicknet on MultiSpert: Fast Parallel Neural Network Training. *ICSI Technical Report TR-97-047.*, http://www.icsi.berkeley.edu/techreports/, 1998.

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", EuroSpeech 1997, pages 1895-1898.