6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

UTTERANCE VERIFICATION USING MODIFIED SEGMENTAL PROBABILITY MODEL

Bin Jia[1], *Xiaoyan Zhu*[2], *Yupin Luo*[1], and *Dongcheng Hu*[1]

[1]Dept. of Automation, Tsinghua University, Beijing 100084, P. R. China
[2]Dept. of Computer Science and Technology, Tsinghua University
zxy-dcs@mail.tsinghua.edu.cn

ABSTRACT

Today speech recognition is requested not only to decode utterances into transcriptions, but also to determine the reliabilities of the result, by Utterance Verification (UV). With the conventional HMM, the measure of reliabilities can not be determined directly by the likelihoods of models. Whereas, Modified Segmental Probability Model (MSPM), suggested in this paper, with its normalized likelihood, facilitates rendering UV and speech recognition at the same time and as a whole. In the paper, Integrated Anti-word Model (IAM) is suggested, which is used to advance the measure of UV likelihood of MSPM. Some experiments show high perfor-mance and moderate computation with IAM.

Keywords: speech recognition, utterance verification, segmental probability model,anti-word model, HMM

1. INTRODUCTION

In the conventional HMM-based speech recognition, the transcription of a certain utterance is gotten by the ML algorithm. The result trainscription is only determined by the relative magnitude of the likelihood scores of all the canditate transcriptions. However, the reliability of the result can not be guaranteed, which is among the factors limiting the broad appliction of speech recognition.

Since Utterance Verification (UV) [1] was suggested, there has been a way to evaluate the likelihood scores of hypothetical transcriptions. So, the validity of recognition can be determined by the verification.

With the conventional HMM, the measurement of the reliabilities of a result transcription can not be achieved straightforwardly by the likelihood of the transcription model for speech recognition. This is because the likelihood of an utterance includes all the probabilities of the frames in the utterance. So the likelihood of an utterance to a HMM model varies greatly with the variation of duration of the utterance. Thus, there can not exist a constant likelihood acting as the threshold for UV.

As is shown in the paper, Modified Segmental Probability Model (MSPM), which is developed from State Duration-based Segmental Probability Model (SDSPM) [2], has the normalized probabity likelihood, which is robust to the variation of utterance duration. So, the likelihood of MSPM can be used directly to Utterance Verification besides speech recognition.

As to enhance the performance of verification, the Integrated Anti-word Model (IAM) is sugested in the measurement of recognition and verification scores. The introduction of IAM, in stead of anti-word model, results in less computation without obvious decrease of performance.

In the following section, we begin with the introduction of MSPM and contrast of its likelihood to HMM's to show that UV and speech recognition can be rendered at the same time in MSPM. Next, Integrated Anti-word Model is introduced in the measurement of verification score. Then, in Section 4, some experiment results with Mandarin isolate word are given. At last, the conclusion is presented.

2. MODIFIED SEGMENTAL PROBABILITY MODEL

MSPM is based on SDSPM [2]. In this section, SDSPM is briefly introduced at first, then MSPM is suggested, together with the contrast to HMM in the term of likelihood.

2.1 SDSPM

The observation sequence of utterance, $O = (O_1, O_2, \Lambda, O_T)$, where T is the number of frames in the utterance, is divided by segmentation algorithm [3], into N segments or states, $S^{(1)}, S^{(2)}, \Lambda, S^{(N)}$, where $S^{(i)}$ comprises the observation vectors sequence $O(i) = (O_{t_{i-1}+1}, \Lambda, O_{t_{i-1}+1})$. The probability of O produced by model λ is

$$f\left(O|\lambda, S^{(1)}, \Lambda, S^{(N)}\right) = \prod_{i=1}^{N}\left(P_i(t_i - t_{i-1}) \prod_{t=t_{i-1}+1}^{t_i} b_i(O_t|\lambda)\right) \quad (1)$$

Where $b_i(O_t|\lambda)$ stands for the Gaussain mixture distribution of O_t at the i-th segment, and $P_i(\bullet)$ for the probability of state duration in the state i. Especially, in the Equation (1),

$$f_i\left(S^{(i)}|\lambda\right) = \prod_{t=t_{i-1}+1}^{t_i} b_i(O_t|\lambda)$$ is the likelihood of observation sequence $O(i)$ in state i.

2.2 MSPM

In SDSPM, the state likelihood of observation sequence, $f_i\left(S^{(i)}|\lambda\right)$, is the multiplication of $b_i(O_t|\lambda)$, the likelihood of observation vector o_t. The likelihood is subjected to the variation of the duration of the segment $S^{(i)}$. In MSPM, $f_i\left(S^{(i)}|\lambda\right)$ is normalized so as to eliminate the effect of the duration of the segment, $S^{(i)}$. One of the simplest normalization methods is the adoption of the weighted geometric mean. That is,

$$\bar{f}_i\left(O^{(i)}, S^{(i)}|\lambda\right) = k_i\left(\prod_{t=t_{i-1}+1}^{t_i} b_i(O_t)\right)^{\frac{1}{\tau_i}} \quad (2)$$

where k_i represents the weight coefficient of geometric mean and $\tau_i = t_i - t_{i-1}$ is the duration. The geometric mean enhances the robustness of likelihood to the variation of duration.

Furthermore, it seems that there exits a vector, O_i', corresponding to the segment $S^{(i)}$, with its likelihood at the segment as follows.

$$f_i\left(O_i', S^{(i)}|\lambda\right) = k_i\left(\prod_{t=t_{i-1}+1}^{t_i} b_i(O_t)\right)^{\frac{1}{\tau_i}} \quad (3)$$

In other word, $O(i) = (O_{t_{i-1}+1}, \Lambda, O_{t_{i-1}+1})$ can be supplanted by the vector O_i' with the equivalent likelihood at segment $S^{(i)}$.

For the whole utterance, the probability likelihood can be reprented as follows.

$$f\left(O|\lambda, S^{(1)}, \Lambda, S^{(N)}\right) = \prod_{i=1}^{N}\left(P_i(t_i - t_{i-1}) f_i\left(O_i', S^{(i)}|\lambda\right)\right) \quad (4)$$

Without the consideration of the probability of the state duration, $P_i(t_i - t_{i-1})$, the likelihood of the utterance is equivalent to the joint probability distribution of the N virtual vectors, O_i', $i=1,\cdots,N$, or in other word, the probability distribution of a fixed-dimention vector, comprising the virtual vectors. So, the variation of the speed of utterance, or the duration of utterance affect little the likelihood of the utterance.

As for HMM, the likelihood of the utterance, under the same segment or state situation, is

$$f\left(O|\lambda, S^{(1)}, \Lambda, S^{(N)}\right) = \prod_{i=1}^{N}\left(\prod_{t=t_{i-1}+1}^{t_i} b_i(O_t|\lambda)\right)\prod_{i=2}^{N}\left(A_{S_{i-1},S_i}\right) \quad (5)$$

where A_{S_{i-1},S_i} is the probability of state transition. From Equation (5), it is obvious that the variation of the duration of each state or the whole utterance will all affect the likelihood of the utterance.

According to the verification theory [4], given an utterance, *O*, and a transcription, λ, there should be a test statistic, $T(O|\lambda)$, and a corresponding threshold, ω. When $T(O|\lambda) \geq \omega$, the utterance is considered to be the generation of transcription, λ. Otherwise, if $T(O|\lambda) < \omega$, *O* is believed generated by some other transcription. Since the likelihood of an utterance to a HMM model varies greatly with the variation of duration of the utterance, there can not exist a constant likelihood acting as the threshold for UV. Whereas, the probability likelihood of MSPM can be directly used not only as the measure for speech recognition, with ML algorithm, but also as the test statistic, $T(O|\lambda)$. And the acceptation threshold for UV exits and can be determined by experiments.

3. INTEGRETED ANTI-WORD MODEL

In order to enhance the performance of UV, like other verification methods, the anti-word model is introduced into the test statistic, $T(O|\lambda)$, as follows:

$$T(O|\lambda) = f(O|\lambda) / f_{anti}(O|\lambda) \quad (6)$$

where $f_{anti}(O|\lambda)$ stands for the probability likelihood of the anti-word of transcription λ. The introduction of the anti-word increase largely the computation. For example, if there are *W* candidate transcriptions, their will be $2 \times W$ probability likelihoods to be calculated. In order to reduce the computation, the Integrated Anti-word Model (IAM) is introduced. Let $f_{AIM}(O)$ denote the probability likelihood of all the *W* transcriptions. The parameters of IAM is trained by all the utterance corresponding to all the *W* transcriptions. With N transcriptions, λ_i, $i = 1, \Lambda, N$, and the IAM, there can be a shreshold, θ_{reject}. If utterance *O* is not uttered by any one of the transcriptions, λ_i, $i = 1, \Lambda, N$, there should be $f_{IAM}(O) < \theta_{reject}$; whereas, if the utterance is uttered by one of the transcriptions, there should be $f_{IAM}(O) \geq \theta_{reject}$. Furthermore, if λ_k, $k = 1, \Lambda, N$, utters the utterance, their must be

$$f(O|\lambda_i) / f_{IAM}(O) \geq 1 \quad (7)$$

and

$$f(O|\lambda_i) / f_{IAMi}(O) < 1, \; i = 1, \Lambda, N, \; i \neq k \quad (8)$$

Thus, with a certain threshold θ, the test statistic of transcription λ_i can be defined as

$$T(O|\lambda_i) = f(O|\lambda_i) / f_{IAMi}(O) \quad (9)$$

If $T(O|\lambda_i) \geq \theta$, λ_i is the transcription of utterance *O*. And if $T(O|\lambda_i) < \theta$, λ_i is not the transcription of the utterance.

Therefore, the joint algorithm of speech recognition and UV is as follows:

Step 1 With $f_{IAMi}(O)$ and θ_{reject}, detect whether the utterance is uttered by one of the *N* transcription.

Step 2 If the transcription of the utterance is among the *N* transcription, determine the proper transcription using $T(O|\lambda_i)$, $i = 1, \Lambda, N$, and θ.

4. EXPERIMENTS

A Chinese isolate word recognition task, in which 100 words is designated as keywords, is chosen for performance evaluation purposes. The database used consisted of a training set collected from 30 adult man, who uttered the 100 keywords and 20 non-keywords, and a testing set collected from other 10 adult man, who uttered the 100 key-words and other 20 non-keywords. The feature vector in experiments consisted of the following 33 parameters: 16 LPC derived cepstral coefficients, 16 delta cepstral coefficients, and normalized delta log energy. There are 3 groups of controlled

experiments as follows: first, the test statistic function, $T(O|\lambda_i)$, is only the probability likelihood of MSPM; second, the test statistic function, $T(O|\lambda_i)$, include the probability likelihood and the corresponding antiword model, which is trained with the words other than the corresponding word; third, IAM is included instead of anti-word model, as $T(O|\lambda_i) = f(O|\lambda_i)/f_{IAM}(O)$. MSPM used in the experiments is of 5 segments (or states) and 32 mixtures per state. The results with 5% reject rate are illustrated in Table 1.

$T(O\|\lambda_i)$	$f(O\|\lambda_i)$	$f(O\|\lambda_i)/f_{anti}(O\|\lambda_i)$	$f(O\|\lambda_i)/f_{IAM}(O)$
Accuracy when accept	97%	99.5%	99%
Normalized computation cost	1	2.01	1.12

Table 1. Results of the three experiments

From Table 1 above that considering the reliable recognition accuracy, the algorithm with IAM is a little worse than the conventional anti-word algorithm. This is because the IAM is trained by all the data of keywords and is not so accurate as the anti-words in the conventional anti-word model, which is just corresponding to a certain keyword. However, the AIM algorithm outperforms the first algorithm with only the probability likelihood of MSPM by 83%. From the computation cost, the second is much larger than the rest two, and the third one is a little higher than the first one. So, the last one is a good solution when we consider both perfomance and computation.

5. CONCLUSION

The MSPM is used in the Utterance Verification. Since it has the normalized likelihood score, Speech Recognition and Utterance Verification can be rendered at the same time. In order to enhance the performance, we introduce the Integrated Anti-word Model, which is used in the measure of verification score. Primitive experiments in isolated word recognition and verification have been carried on, which show high performance and small computation cost.

6. REFERENCES

[1] Rahim M.G., Lee C.H. and Juang B.H., Robust Utterance Verification for Connected Digits Recognition. *ICASSP'95*, v.1-5, pp.285-288, 1995

[2] Jia B., Zhu X.Y., Luo Y.P. and Hu D.C., State Duration-based Segmental Probability Model. *ICCT'98*, v.2, ps041-5, 1998

[3] Huang C, *Language information processing*, Beijing: Tsinghua University Press, 1996

[4] Rahim M. G., Lee C. H., And Juang B. H., Discriminative Utterance Verification for Connected Digits Recogniton. *IEEE Trans. On Speech and Audio Processing*, Vol.5, No.3, p266-277, 1997