EVALUATION AND IMPLEMENTATION OF A VOICE-ACTIVATED DIALING SYSTEM WITH UTTERANCE VERIFICATION

Beng T Tan, Yong Gu, and Trevor Thomas

VOCALIS Ltd.
Chaston House, Mill Court, Great Shelford,
Cambridge CB2 5LD, UK
E-mail: beng@vocalis.com

ABSTRACT

This study investigates the utterance verification (UV) algorithm for a voice-activated dialing (VAD) system. The UV techniques help to improve the system accuracy of a VAD system and to improve the efficiency of user interface by reducing the need of confirmation. In this paper, we examine various UV methods, namely, all-phone garbage model (GM), N-best likelihood ratio (NBLR), and the combined methods. The performances of a VAD system with UV at various vocabulary sizes are studied. By rejecting 9.5% of correctly recognized names, the system error rate become less than 0.3% which represent a reduction of 91% in error rate over the baseline system. The UV technique can reduce the number of confirmation by at least 88% with a system error rate of 0.28%.

1. INTRODUCTION

There is an increase in demand in the marketplace for a voice-activated data retrieval system that uses audio to enter and retrieve data. One of the important applications is voice-activated dialling (VAD) system, which allows the users to dial a number by simply saying the name [1,6].

It is very likely that the user will enrol an entry into the system using real names or user-specific jargon. A VAD system should be able to cope with an open and fully customised vocabulary set and therefore a task specific vocabulary can not be pre-defined. It differs from conventional speech recognition system in the way that the pronunciation of a keyword has to be transcribed by the machine.

Confirmation is generally required after recognition to ensure that the correct number is dialled. A user-friendly VAD system will minimize the need of confirmation. This can be accomplished through utterance verification (UV) which is a process to verify the keyword hypotheses produced by a speech recognizer.

The UV method based on garbage models has been applied to VAD with the purpose to reject out-of-vocabulary (OOV) [2]. Rejection of OOV is not our primary concern because they happen less frequent in a speaker-dependent system like VAD. We are more interested in improving the system performance and improving the user interface by reducing the need of confirmation. The purposes of applying UV to VAD are to categorise the results into correct recognition, uncertain recognition, incorrect recognition, and OOV. Confirmations are only required when the recognition results are uncertain. The system rejects the incorrect recognition and OOV events. In a VAD application, certain level of rejection rate can be tolerated as the caller can easily try again. In this case, the overall system performance is enhanced. In this paper, we examine various UV methods, namely, all-phone garbage model (GM), N-best likelihood ratio (NBLR), and the combined methods.

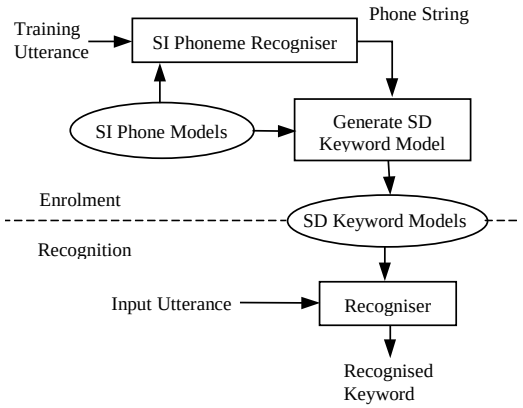

Figure 1: Voice-activated dialling system

2. VOICE-ACTIVATED DIALLING SYSTEM

The overall architecture of a VAD system is shown in Figure 1. An isolated example of each keyword is presented to the system during the enrolment session. A speaker-independent (SI) phone recogniser is used to transcribe the new string using tri-phone models which are sex-segregated. A sequence of phone labels is generated from the phone recogniser. The phone string generated by the phone recognizer during the enrolment session is used in the grammar during the recognition phase. Since the pronunciation is derived from a single speaker, the word models represented in the grammar are speaker-dependent. However, if the pronunciations of different speakers are fairly close to each

other, the same system can also be used in a multiple user environment [6].

The SI phone models are modelled by continuous density multiple Gaussian distributions hidden Markov model (HMM). Each model has 3 emitting states, with a left-to-right topology, and 10 mixture components. The parameters used in the system included 12 cepstral coefficients and 12 delta cepstral coefficients giving a vector size of 24.

3. UV CONFIDENCE SCORES

Utterance verification is a hypothesis-testing problem. The aim of hypothesis testing is to decide whether to accept a null hypothesis, H_0, or to accept an alternative hypothesis, H_1. We need a test statistic to carry out the hypothesis decision. The likelihood, $P(O|h)$, generated by a conventional HMM-based recogniser cannot be used directly for this purpose because it is relative to a particular acoustic observation, O, and it is not comparable across utterances.

A popular UV score is the N-best likelihood ratio (NBLR) [3]. The verification score for hypothesis k is the log of the likelihood ratio of candidates k and $k+1$ and is represented by

$$R(O,k) = \frac{\log(P(O \mid h_k)) - \log(P(O \mid h_{k+1}))}{\text{number of frames}}, \qquad (1)$$

where h_k and h_{k+1} is the kth and $k+1$st hypotheses produced by an N-best algorithm, respectively. This measure gives us the exact ratio of $P(h_k|O)$ and $P(h_{k+1}|O)$. The merit of this method is that it requires no additional models to perform UV. This technique is particular good for detecting the mis-recognized events. Ideally, the score $P(O|h_1)$ should be much higher than the score $P(O|h_2)$ if h_1 is the correct hypothesis.

Alternatively, the score of the top candidate in the N-best list can be normalised by a garbage model (GM) using free grammar phone models running parallel with the keyword models [4],

$$G(O) = \frac{\log(P(O \mid h_1)) - \log(P(O \mid h_g))}{\text{number of frames}}, \qquad (2)$$

where h_g is the all-phone garbage models. The advantage of this garbage model is that no specific garbage samples are required for training. This method is particular useful for rejection of OOV words. In this paper, we investigate two garbage models, one consists of context independent phones (G_{ci}) and another consists of triphones (G_{tris}).

Since difference confidence scores are derived from different set of parameters, they may be combined into a new confidence score that can outperform either measure when considered separately. The UV score can be combined as a linear combination of two likelihood scores [5]. Alternatively, we apply a strict rule to various scores for hypothesis testing. In a one-tailed test, the critical region is defined by a single threshold value. When various critical regions are considered, a hypothesis is accepted only if it falls within the overlapping area of these critical regions. This is expressed as follows:

$$C(R,G) = \begin{cases} 1, & \text{if } (R(O,k) > T_r \text{ and } G(O) > T_g), \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

where T_r and T_g are the thresholds for $R(O,k)$ and $G(O)$. In the current experiment, these two thresholds are set such that the rejection rates of correctly recognised tokens are the same when they are applied separately.

4. EXPERIMENTS

4.1 Baseline Performance

The database used in our evaluation consists of speech collected in Vocalis from 28 speakers. Each speaker uttered 40 name strings and recorded each name with three repetitions. There are another 10 names from each speaker reserved for evaluating OOV performance. Every speaker pronounces a different set of names. The phone models are trained on a separate database.

We vary the vocabulary size from 2 to 40 and the baseline performances are shown in Figure 2. The error rate is almost linearly proportional to the vocabulary sizes. The difference in error rate between vocabulary size of 2 and 40 is about 2.4%.

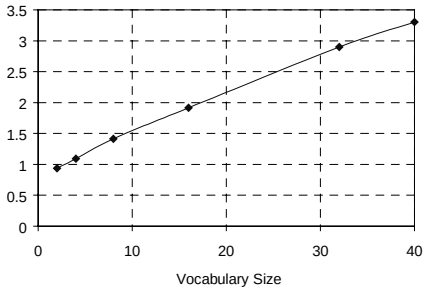

Figure 2: Baseline performance of VAD at various vocabulary sizes.

4.2 Comparative Results of UV Methods

Figures 3 and 4 compare the performances of the UV scores $G_{ci}(O)$, $G_{tris}(O)$, $R(O,1)$ and combined confidence scores $C(R,G)$ at vocabulary size of 40. Figure 3 shows the system error rate after UV versus the false rejection rate of names that are correctly recognised. If the system can tolerate 5% of the correctly recognised names being rejected, the confidence scores $R(O,1)$, $C(R,G_{ci})$, and $C(R,G_{tris})$ can reduce the system error rates of the baseline system by 79%. The NBLR method and combined methods are better than garbage models for rejecting mis-recognized name strings. The garbage model using context-independent phones perform slightly better than the garbage models using triphones.

Figure 4 shows curves of the rejection rate of OOV names versus the false rejection rate of correctly recognized names. At

an operating point of 5% rejection of correctly recognized names, the garbage methods and the combined methods can reject more than 85% of the OOV name strings, while the performance of the NBLR method is significantly lower (about 77%).

It is evident from Figure 3 and 4 that the garbage models and NBLR methods perform quite different in rejecting mis-recognized name and OOV events. Since the merits of these two UV methods are distinct, by applying the strict rule to combine them as in Eqn (3) we can preserve their individual merits and achieve the best performance.

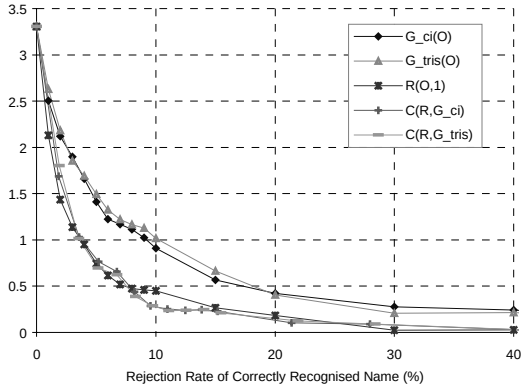

Figure 3: UV performance on mis-recognized valid names at vocabulary size of 40.

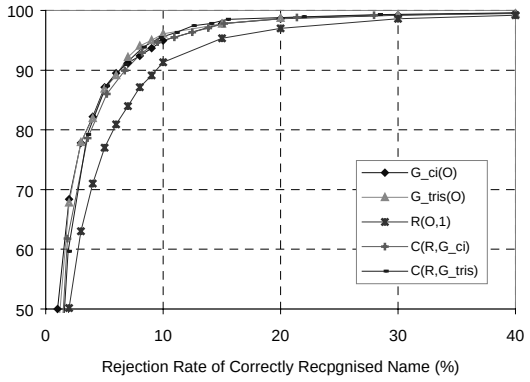

Figure 4: UV performance on OOV name strings at vocabulary of 40.

To compare the effect of vocabulary size on various UV methods, we vary the vocabulary size from 2 to 40. Figure 5 shows the rejection rates of correctly recognized name strings as a function of vocabulary size. The threshold is chosen at vocabulary size of 40 such that all UV methods have a rejection rate of about 10%. As the vocabulary size changes from 2 to 40, the variation in rejection rate is about 3% for the combined method compared to 8.4% for the NBLR method and 4% for the garbage model methods. The combined methods perform most even at various vocabulary sizes.

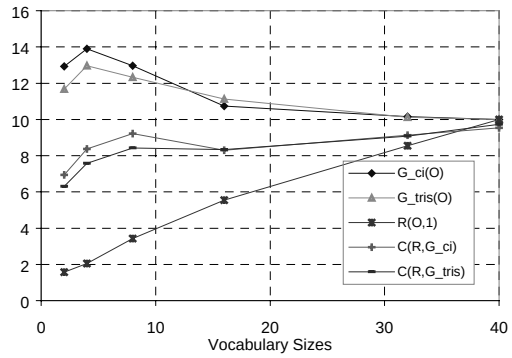

Figure 5: Rejection rate of correctly recognized names versus vocabulary sizes.

Figure 5 also suggests that the garbage model is less dependent on the vocabulary size than the NBLR method. This is evident in Figure 6 and 7 which show how the thresholds of the score $G_{tris}(O)$ and the score $R(O,1)$ vary as the vocabulary sizes increase. .

Every curve in Figure 6 and 7 represents the threshold values versus the vocabulary sizes at a particular rejection rate of correctly recognized names. Rejection rates ranging from 5% to 50% are shown in the figures.

To maintain at a constant rejection rate at various vocabulary sizes, the thresholds of the score $G_{tris}(O)$ remain fairly constant as comparing to the threshold of the score $R(O,1)$. One explanation for the strong dependence of the threshold of the score $R(O,1)$ on vocabulary size is the fact that the score of the second best hypothesis depend on the vocabulary set.

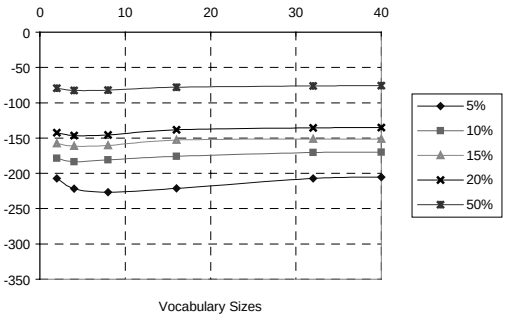

Figure 6: Threshold of the $G_{tris}(O)$ at various vocabulary sizes.

4.3 Performance of Combined Method

From the results in the previous sections, it is clear that the combined methods are the best in term of the system error rate, the OOV rejection, and the stability of rejection rate of correctly recognized name at various vocabulary sizes. In this section, we focus on evaluating the performance of the combined method $C(R,G_{ci})$.

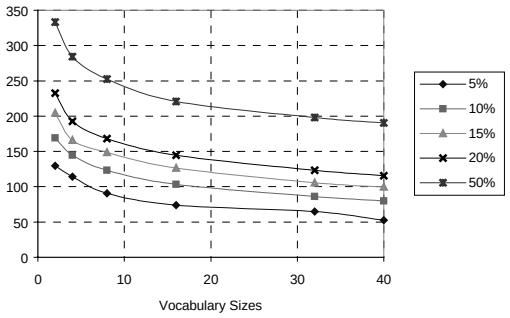

Figure 7: Thresholds of the $R(O,1)$ at various vocabulary sizes

The threshold is set at vocabulary size of 40 with a 9.5% rejection rate of correctly recognized name strings and the vocabulary sizes vary from 2 to 40. The system error rates and OOV rejection rates at various vocabulary sizes are shown in Table 1. The system error rate is maintained below 0.3% and more than 94% of OOV events are rejected at various vocabulary sizes.

Vocabulary Size	2	4	8	16	32	40
System Error (%)	0	0	0.09	0.13	0.18	0.28
OOV Rejection (%)	99.6	99.4	98.5	97	94.8	94.7

Table 1: UV performance of the combined method, $C(R,G_{ci})$, with a 9.5% rejection rate of correctly recognised name strings.

Suppose the vocabulary size is 40 and there is no confirmation after each call, the system error rate is 3.3%. When there is no UV and each call requires a confirmation, the system error rate is ideally 0% and the confirmation rate is 100%.

If we apply the UV technique and choose an operating point with a rejection rate of correctly recognised name strings of 9.5%, the system error rate becomes 0.28% (Table 1). The overall rejection rate (regardless correctly or incorrectly recognized name) is 12%. The system can either reject all of these entries i.e. confirmation rate is 0% or confirm all of these entry i.e. confirmation rate is 12%. If a second threshold is set such that the poorly recognized entries will be rejected straight away, the confirmation rate can range from 0% to 12%. In other words, we can at least reduce the confirmation rate by 88% with a system error rate of 0.28%.

5. CONCLUSIONS

In this paper, we presented a voice-activated dialling (VAD) systems with utterance verification (UV). We study three UV methods, namely, an all phone garbage models (GM), the N-best likelihood ratio (NBLR), and the combination of these two. The combined approach was shown to outperform either method when considered separately. The NBLR score is useful for detecting mis-recognized events and the all-phone GM score is useful for rejecting out-of-vocabulary (OOV) events. The combined method has the merits of both.

It is found that the threshold setting of GM is less dependent on vocabulary size than the NBLR method. The performances of the combined method are more even across different vocabulary sizes.

By rejecting 9.5% of correctly recognised names, we can achieve a system recognition rate of more than 99.7% and rejecting more than 94% of OOV events. When there is no UV, the difference in error rate between vocabulary size of 2 and 40 is about 2.4%. This difference is reduced to 0.28% when UV is applied. The UV helps to limit the system error rate as the vocabulary size increases. We can also improve the user interface by reducing the number of confirmation through UV by at least 88%.

6. REFERENCES

1. Buhrke, E.R., Jacobs, T., and Wilpon, J. "A Comparison of Algorithms for Speaker-Dependent Speech Recognition for Network Applications," *Proc. IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, 1992.

2. Ramalingam, C.S., Netsch, L., and Kao, Y.-H. "Speaker-Independent Name Dialing with Out-of-Vocabulary Rejection," *Proc. ICASSP*, 1997.

3. Caminero-Gil, F.J., de la Torre, C., Hernandez-Gomez, L.A., and Martin-del Alamo, C. "New N-best Based Rejection Techniques for Improving a Real-Time Telephonic Connected Word Recognition System," *Proc. Eurospeech*, Madrid, Vol. 3, pp.2099-2102, 1995.

4. Young, S.R., and Ward W. "Recognition Confidence Measures for Spontaneous Spoken Dialog," *Proc. Eurospeech*, pp. 1177-1179, 1993.

5. Sukkar, R.A., Setlur, A.R., Lee, C.H., and Jacob, J. "Verifying and Correcting Recognition String Hypotheses Using Discriminative Utterance Verification," *Speech Communication*, pp. 333-342, 1997.

6. Tan, B.T., Gu, Y., and Thomas, T. "Implementation and Evaluation of a Voice-Activated Dialling System," *Proc. IVTTA*, Torino, Italy, 1998.