HANDS-FREE VOICE ACTIVATION OF PERSONAL COMMUNICATION DEVICES

Sahar E. Bou-Ghazale and Ayman O. Asadi

Conexant Systems, Inc.
Media Access, Speech Technology Development
4311 Jamboree Road, K02-250
Newport Beach, CA 92660, USA
sahar.bou-ghazale@conexant.com *ayman.asadi@conexant.com*

ABSTRACT

We propose a real-time speaker-independent voice activation system composed of an HMM-based continuous speech recognition algorithm and a post-processing component. The post-processing component validates the recognition results by applying a set of criteria and confidence measures. Voice activation in this context refers to a system designed to continuously listen to incoming unconstrained speech and to recognize a pre-defined activation command while rejecting any other extraneous speech or noise. The proposed system is intended for turning a device on or for activating a main speech recognition application upon recognizing the activation utterance. This paper includes a description of the overall system, the post-processing algorithm formulation, and a discussion of the final results. The proposed system has been trained and tested on a database comprised of 2000 North American English speakers. The system yields a 0.55% false alarm rate, and a 15% false rejection rate for both isolated and embedded utterances.

1. INTRODUCTION

Current personal communication systems such as a cellular phone with integrated speech recognition capabilities require the user to manually activate the device or the speech recognition system by pushing a button before the user can issue voice commands to control the device. In order to offer a complete hands-free operation in busy-hands busy-eyes environments, we propose using a voice activation system for turning the device on simply by speaking a pre-selected activation utterance. The voice activation system continuously listens to and processes incoming speech. Upon recognizing that the activation command has been spoken, the voice activation system turns the device on or switches to the main speech recognition system to receive further commands from the user. Such a feature offers a needed convenience to the user especially where total hands-free operation is required such as operating a cellular phone in a moving car environment.

The voice activation system could be thought of as a keyword spotting system that continuously processes incoming speech as opposed to processing limited duration utterances. The system can be realized by training a keyword model for the desired activation word, a garbage model representing both extraneous speech and background noises, and a model for background silence [1][2]. However, several challenges are involved in designing the proposed system that is intended to recognize keywords embedded in running speech rather than in limited duration utterances. For example, since the user is not required to speak within a specified period of time, the system ought to make a decision on a continuous basis whether the desired word was spoken or not. In addition since the system is running continuously, maintaining low false alarm rates and low-power consumption are critical.

The remainder of this paper is organized as follows. Section 2 includes an overview of the voice activation components. Section 3 includes a detailed discussion of the criteria and flow graph of the post-processing component. In Section 4, we discuss the database employed for training and recognition, as well as the speech activation performance when tested on isolated words as well as on running speech. Final conclusions are given in Section 5.

2. VOICE ACTIVATION SYSTEM COMPONENTS

The voice activation system is comprised of an endpoint detection algorithm, a speech recognition algorithm and a post-processing component. Whenever speech activity is detected, the speech recognition component processes the input speech on a frame-by-frame basis and makes a decision at every frame whether the activation word has been spoken. The post-processing component validates the speech recognition result. The components of the voice activation system are discussed in the following sections.

2.1 Endpoint Detection

The end-point detector employed here is an energy-based algorithm. The end-points are extended on both sides by a small number of speech frames to ensure that the actual end-points are within the utterance. Note that the speech recognition system does not require explicit endpoint detection of speech since the system is designed to handle unconstrained sequence of silence, noise, and extraneous speech. However, by incorporating an endpoint detector, the recognizer will start processing only upon detection of speech which reduces power consumption, and as a result the false alarm rate drops by a factor of 2.

0-7803-6293-4/00/$10.00 ©2000 IEEE.

Authorized licensed use limited to: IEEE - Staff. Downloaded on January 06,2025 at 20:24:00 UTC from IEEE Xplore. Restrictions apply.

2.2 Real-Time Speech Recognition System

The recognition system is frame-synchronous and is based on continuous density hidden Markov models (CD-HMM). All the models are trained using a mixture of Gaussian distributions to represent the output probabilities [3][4]. The recognition system allows for silence or extraneous speech to be spoken before or after the activation phrase. In other words, the recognizer can detect the word whether it is spoken in isolation or embedded in extraneous speech. The input speech is sampled at 8kHz, then filtered by a band-pass filter with a pass-band from 300 to 3300 Hz. The filtered speech is processed at 20 ms frame rate with a Hamming window of 30 ms. A total of 10 LPC-based cepstral parameters, 10 delta-cepstra and 1 delta-energy are computed for each frame.

The recognition system uses a finite state network, as shown in Figure 1, to represent the allowable sequences of silence, keyword, and out-of-vocabulary (OOV) words. The finite state network consists of optional silence and OOV models at the beginning and end, and a parallel network of the activation keyword and one OOV model. The operating point of the system can be adjusted by varying the weight or probability associated with the keyword and OOV model. The operating point is chosen to optimize the false rejection and false alarm rates. The weight associated with the optional OOV model (at the beginning and end) is different than that associated with the OOV model in parallel with the keyword.

The following are some examples of allowable sequences:

- [Silence Keyword Silence]
- [Silence OOV Silence]
- [Silence Keyword OOV Silence]
- [Silence OOV Keyword Silence]
- [Silence OOV Keyword OOV Silence].

In these examples, the silence and OOV can be repeated many times to represent all extraneous silence and speech before or after the keyword, or the utterances that do not include the keyword.

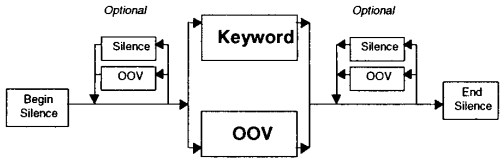


Figure 1. A finite state network representing the allowable word sequences.

The recognition algorithm performs a trace-back at every frame, and passes the result string with the highest likelihood score to the post-processing component.

2.3 Post-Processing Component

The post-processing component decodes the result string on a frame-synchronous basis and applies a set of criteria and confidence measures to decide whether the activation word was a part of the utterance. The activation word is declared to be found only after it meets a set of statistics derived from training data such as the overall duration of the keyword, the number of times the keyword appears in the result string of consecutive frames, and the confidence level associated with the keyword. The steps of the algorithm are illustrated in the block diagram of Figure 2 and are discussed in more detail in the following section.

3. FORMULATION OF THE POST-PROCESSING ALGORITHM

The following set of criteria is developed and later integrated with the post-processing algorithm to make the final decision.

3.1 List of Criteria:

Criterion 1: The recognized utterance should satisfy a minimum word duration, α, to be determined from the training data.

Criterion 2: The recognized utterance should satisfy a maximum word duration, β, to be determined from the training data.

Criterion 3: The activation word should appear in the result string of consecutive frames for a minimum number of frames, γ, to be determined experimentally.

Criterion 4: The activation word can appear in the result string of consecutive frames for a maximum number of frames, ψ, after which we check the remaining criteria.

Criterion 5: If the activation word has not appeared in the result string of consecutive frames within the past T seconds, then reset the parameters and start the search over again. If the keyword has appeared consecutively in the result string, the search is continued past T seconds. This criterion is introduced to reduce the memory storage for the trace-back without compromising the performance.

Criterion 6: The relative log-likelihood score should be equal to or greater than a threshold, η, given by:

$$\eta = (\text{Score}_{\text{keyword}} - \text{Score}_{\text{OOV}}) / |\text{Score}_{\text{keyword}}|,$$

where the accumulated log-likelihood score for the keyword model is denoted by $\text{Score}_{\text{keyword}}$, and the log-likelihood score for the OOV model is denoted by $\text{Score}_{\text{OOV}}$.

3.2 Post-Processing Algorithm Flow Graph

At this point, we formulate the post-processing algorithm based on the set of criteria mentioned previously. The steps of the algorithm are illustrated in the block diagram of Figure 2. Block **A** verifies that the first and second criteria have been met. Once the keyword duration requirement has been satisfied and the keyword has appeared in the result string of consecutive frames, block **D** verifies that the third and fourth criteria have been met. At this point, the search continues until (a) the detection of end of speech, (b) criterion 4 is met, or (c) the keyword stops appearing in the result string. Blocks **H** and **M** verify that the third criterion has been met. Block **I** is associated with criterion 5, and finally block **E** is associated with the criterion 6.

There are three paths in the flow diagram of Figure 2 that meet all the requirements and hence result in activating the main speech recognition system. The remaining side paths will either

Authorized licensed use limited to: IEEE - Staff. Downloaded on January 06,2025 at 20:24:00 UTC from IEEE Xplore. Restrictions apply.

terminate the search and start over again, or will continue the search since a theory is developing. A number of these possible paths are given below.

In the path traversing **A, B, C, D, E, F,** the keyword has met all the necessary requirements, and hence the post-processor declares that the *activation word has been found.*

In the path **A, B, C, D, L, M, E, F,** end of speech has been reached and the keyword has met all the necessary requirements. Hence, the post-processor declares that the *activation word has been found.*

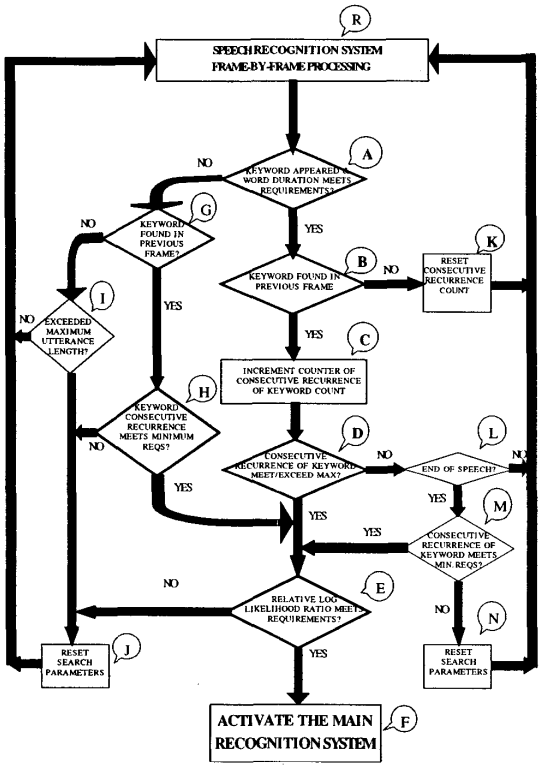


Figure 2.Flow graph of the post-processing algorithm

In the path of **A, G, H, J, R,** the keyword has met criterion 1 and 2 but has failed criterion 3. This indicates that the keyword has appeared in the result string for few consecutive frames, but has eventually dropped out of the results before satisfying the third criterion. The following section summarizes the database employed in our evaluations.

4. EVALUATIONS

4.1 Database

The recognition system was trained on one database and tested on two different databases. The database used for training was intended for command and control applications for personal communications devices. The utterances were mostly isolated keywords and a small portion was embedded keywords in phrases, background speech and other non-speech sounds. The database was collected from 1000 speakers over the telephone network and 1000 in-person speakers (i.e., without the telephone network impairments) using various handset types. The speakers were distributed over various genders, age groups, dialects of North American English and various venues where a communication device could be located. The speakers recorded the utterances in two different recording sessions. The handset type was changed between the two recording sessions for the in-person speakers. The over the phone speakers called again our data collection system and that could have been routed through a different part of the PSTN. A total of 153, 096 utterances from this database were employed for training, while the remaining 19,423 utterances were set aside for testing.

A second in-person database consisting of 50 sentences recorded by 27 male and female speakers using a microphone was collected in-house for further evaluations. Examples of utterances collected are "Set-up my wake-up call for 7 AM", "Wake-up the voice activation program", "Fax a copy of my journal to Bob", "Send this memo to the program manager", etc. In the following evaluations, the specific activation utterance is chosen as the word "wake-up".

4.2 HMM Model Training

The training data was labeled as either voice activation word or extraneous speech. A total of 2300 utterances were employed for training the activation word model. For the out-of-vocabulary model training, all the extraneous speech regions, noises and 140,796 non-keyword utterances were combined together to train one HMM model. A separate 3-state model was trained for background silence. In the evaluations presented here, the number of states for the keyword and OOV models was set to 10, and the number of mixtures per state was set to 3. Several other values were evaluated for the number of HMM states and mixtures per state. However, we did not observe significant improvements in performance beyond 3 mixtures per state. In addition, the 8 and 10 HMM-states OOV achieved similar performance, while the 1-state OOV resulted in higher false rejection rates.

4.3 Evaluation Data

The test data containing the activation word "wake-up" consisted of 293 isolated utterances and 47 sentences. The test data containing out-of-vocabulary words consisted of 19,130 isolated utterances and 1089 sentences. The evaluations were all conducted using simulated running speech since long sessions of actual running speech containing the activation utterance were not readily available. Running speech was simulated by continuously reading from speech files as if they were one continuous utterance. This resulted in 28.5 hours of speech.

4.4 Experimental Results

In this section, we discuss a set of experiments that were conducted to determine the set of parameters T, α, β, γ, ψ, and η for the activation word "wake-up". We also discuss the voice activation system performance on isolated and embedded speech.

Authorized licensed use limited to: IEEE - Staff. Downloaded on January 06,2025 at 20:24:00 UTC from IEEE Xplore. Restrictions apply.

In order to reduce the memory storage required for trace-back without compromising the performance, the shortest search duration, T, was experimentally set to 3.5 seconds. By clearing the history and starting the search over again if the keyword was not detected within 3.5 seconds, the false rejection rate was reduced by a factor of 13.

For deriving the word duration statistics, all of the training data for the activation word was employed. The keyword duration for the activation word "wake-up" varied from 0.18 seconds to 1.65 seconds, when spoken in isolation, with an average duration of 0.6 seconds. In the following recognition evaluations, the minimum word duration requirement, α, was set at 0.3 seconds, and the maximum word duration requirement, β, was set at 1 second.

The required number of minimum consecutive frames, γ, was determined by evaluating the recognition performance as a function of γ. The false rejection and false alarm rates are shown in Table 1 for various values of γ. As the value for γ is increased, the false rejection rate increases while the false alarm rate decreases. By setting γ to 10 frames, the false rejection rate increases only by a factor of 2 while the false alarm rate drops by a factor of 8, as compared to $\gamma = 1$. In the following recognition evaluations, γ was set at 10, and the maximum number of allowable consecutive frames, ψ, was set at 50 frames.

Based on our findings, we set-up a voice activation system where $\alpha = 0.3$ sec, $\beta = 1$ sec, $T = 3.5$ sec, $\gamma = 10$ frames, and $\psi = 50$ frames. The system performance on isolated and embedded speech is summarized in Tables 2 and 3 for various threshold values. The number of utterances utilized in each recognition test is also included since it varied from one test to another. It can be seen that without adding a threshold criterion (i.e., $\eta = 0$), the system achieves close to 15% false rejection and 0.55% false alarm rates for both isolated and embedded utterances. As the threshold value is increased, the false rejection rate increases and the false alarm rate decreases. The desired operating point can be determined according to the specific application. If lower false rejection rates are desired, i.e., below the values tabulated in the results, at the expense of higher false alarm rates, then the weight associated with the OOV model that is in parallel with the keyword model can be decreased as discussed earlier in Section 2.2. For achieving very low false alarm rates of 0.26% for isolated utterances and 0% for embedded utterances, η can be set to 0.007.

5. CONCLUSION

In this paper, we have presented a real-time voice activation system for activating a device or a speech recognition system simply by speaking a pre-selected utterance. The voice activation system achieves 0.55% false alarm rate and 15% false rejection rate for both isolated and embedded utterances, and is currently running in real-time in an office environment. The system requires relatively low power and can be implemented in a personal communication device due to its low memory and MIPS requirements (< 1 MIPS). In addition, the voice activation system can be designed in such a way that once the activation word is recognized, the device turns on or switches to the main recognition system and waits to be prompted by a word from the main menu of the speech recognition system. If no valid word is spoken in the following 10-15 seconds, then the system switches back to the stand-by mode and waits to be activated again. Hence in a false alarm situation, the user would not notice that the system got activated since there is no prompt that's being played. The system simply switches to the recognition mode for a short period of time.

ISOLATED WORDS	MINIMUM CONSECUTIVE FRAMES γ			
	1	5	10	15
False Rejection	7.50%	13%	15%	25%
False Alarm	4.30%	1%	0.55%	0.36%

Table 1 . False rejection and false alarm rates of the voice activation system as a function of the minimum number of consecutive frames γ ($\alpha = 0.3$ sec, $\beta = 1$ sec, $T = 3.5$ sec).

ISOLATED WORDS	Relative Log-Likelihood threshold		
	$\eta = 0$	$\eta = 0.007$	$\eta = 0.019$
False Rejection (293 utterances)	15%	17%	28%
False Alarm (19130 utterances)	0.55%	0.26%	0.08%

Table 2. False rejection and false alarm rates of the voice activation system as a function of η when tested on isolated speech ($\alpha = 0.3$ sec, $\beta = 1$ sec, $T = 3.5$ sec, $\gamma = 10$ frames).

EMBEDDED UTTERANCE	Relative Log-Likelihood threshold		
	$\eta = 0$	$\eta = 0.005$	$\eta = 0.007$
False Rejection (47 utterances)	15%	25%	31%
False Alarm (1089 utterances)	0.64%	0.18%	0%

Table 3 False rejection and false alarm rates for the voice activation system as a function of η when tested on embedded speech ($\alpha = 0.3$ sec, $\beta = 1$ sec, $T = 3.5$ sec, $\gamma = 10$ frames).

6. REFERENCES

[1] Wilpon, J. G., Rabiner, L. R., Lee, C-H., Goldman, E. R., "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. On ASSP*, Vol. 38, No.11, pp. 1870 – 1878, Nov. 1990.

[2] Wilpon, J. G., Miller, L. G., Modi P., "Improvements and applications for key word recognition using hidden Markov modeling techniques", *Proc. ICASSP*, pp. 309 – 312, 1991.

[3] Asadi, A., Lubensky, D., Madhavrao, L., Naik J., Raman, V., Vysotsky G., "Combining Speech Algorithms Into A Natural Application Of Speech Technology For Telephone Network Services", *EUROSPEECH*, vol. 1, pp. 273 – 276, September 1995.

[4] L. R. Bahl, F. Jelinek, and R. Mercer. "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 2, pp. 179-190, March 1983.

Authorized licensed use limited to: IEEE - Staff. Downloaded on January 06,2025 at 20:24:00 UTC from IEEE Xplore. Restrictions apply.