ROBUST END-OF-UTTERANCE DETECTION FOR REAL-TIME SPEECH RECOGNITION APPLICATIONS

Ramalingam Hariharan, Juha Häkkinen, Kari Laurila
Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
Email: {ramalingam.hariharan, juha.m.hakkinen, kari.laurila}@nokia.com

ABSTRACT

In this paper we propose a sub-band energy based end-of-utterance algorithm that is capable of detecting the time instant when the user has stopped speaking. The proposed algorithm finds the time instant at which many enough sub-band spectral energy trajectories fall and stay for a pre-defined fixed time below adaptive thresholds, i.e. a non-speech period is detected after the end of the utterance. With the proposed algorithm a practical speech recognition system can give timely feedback for the user, thereby making the behaviour of the speech recognition system more predictable and similar across different usage environments and noise conditions. The proposed algorithm is shown to be more accurate and noise robust than the previously proposed approaches. Experiments with both isolated command word recognition and continuous digit recognition in various noise conditions verify the viability of the proposed approach with an average proper end-of-utterance detection rate of around 94% in both cases, representing 43% error rate reduction over the most competitive previously published method.

1. INTRODUCTION

Despite recent improvements in the automatic speech recognition (ASR) technology, there are still a number of challenges to be solved before ASR can be successfully utilised in various real world environments by novice users. In 1997, Sagayama conducted an internet-based discussion of what is hindering the wider use of ASR [1]. The leading reasons found in the survey were:

- Robustness against noise, utterance and channel variabilities is too poor
- The human interface is too poor for users to utilise ASR comfortably

One of the biggest reasons for the inadequacy of the human interface is the perceived unpredictability of ASR. Clearly the fact that speech recognisers make more errors than humans is the biggest reason for unpredictability. Another important aspect in the perceived predictability is the timing of the ASR system's response to the user. Sometimes the ASR system responds too early, interrupting the user's speech, and sometimes user needs to wait too long for a response. This kind of behaviour greatly annoys users and thus, there is a clear need for ASR systems whose timing of the reply for the user is more predictable, and furthermore, remains similar across wide range of usage environments and noise conditions.

In order to give timely responses to the user, an ASR system should be able to detect when the user has stopped speaking. In our terminology, an end-of-utterance detection algorithm should be utilised. The state-of-the-art end-of-utterance detectors typically utilise methods based on the use of spectral full-band energy values, zero-crossing detectors etc. [2]. However, as shown also in this paper, they do not seem to work well in noisy conditions. Another conventional method to detect the end-of-utterance (EOU) is to use the confidence score of the recognised utterance [3][4][5]. However, the confidence based EOU scheme also works poorly in noisy environments. The main reason is that acoustic models generally do not match noisy signals well, thus producing lower confidences. The spoken words may then not be detected, producing a long delay in the response to the user, especially in noisy conditions. The confidence measure is also likely to produce too-early detections[1] in a recognition system, where the speech vocabulary contains words in which one utterance is part of another longer utterance. Thus, for example, if there are utterances like "*call*" and "*call home*" and the utterance "*call home*" is spoken, it is likely that "*call*" is recognised since it produces a very high confidence first.

In this paper, we describe a sub-band based EOU scheme, which works well in various noise environments. The end-of-utterance detector is developed for both isolated and continuous word recognition applications. The detection is based on finding the time instant at which many enough sub-band energies fall below adaptive thresholds, in other words a non-speech region is detected after the utterance has been spoken. It is verified through experiments that better noise robustness can be obtained by combining the independent decisions of the different sub-bands. The description of the EOU scheme and the associated experimental details for isolated word recognition are explained in the next two sections. Section 4 describes the adaptation of the algorithm for connected word recognition. Finally the conclusions are drawn in the last section.

2. END-OF-UTTERANCE FOR ISOLATED WORD RECOGNITION

The end-of-utterance algorithm for isolated word recognition tries to estimate the time instant when the user has stopped speaking and remained quiet for a pre-defined period of time.

[1] An utterance is said to be detected too-early, when the system stops recognition before the end of the spoken utterance. Hence only part of the utterance is actually recognised.

This time instant is found by examining the sub-band energy trajectories of the utterance. The algorithm uses M sub-band energies, obtained from the mel-scaled energy values, to detect the end of utterance. The log sub-band energies are then used as input to a rank order filtering block, described later in this section, to find the end of utterance hypothesis for each sub-band separately. If many enough individual sub-bands (*n* out of *M*) trigger their sub-band end-of-utterance hypothesis, then the end-of-utterance is said to be detected and recognition is stopped.

The sub-band energy trajectories, obtained from speech, contain many impulse type noises especially in adverse conditions. Since it is well known that non-linear filtering techniques can be employed to reduce the effect of impulsive noise, we did a series of experiments and developed a rank-order filtering scheme, which is described in detail below.

The end-of-utterance is detected separately in each sub-band according to a rank order filtering based EOU scheme, which is described using a state machine representation as shown in Figure 1. The sub-band energy at time *t* is denoted as *p(t)*. In the beginning of the recognition (state *S*0 in Figure 1), initialisation *init()* is done. This means that the counter *c* is set to zero, the global minimum power *p_min(t=1)* is set to ∞ and the global maximum power *p_max(t=1)* is set to $-\infty$. This is followed by an instant transition into state *S*1. The function *f()* in states *S*1-*S*3 consists of the following actions:

1) p(*t*) is buffered (FIFO of *N* last frame energies is used)
2) the global minimum and maximum powers (*p_min(t)* and *p_max(t)*) are computed from a sliding buffer as given by the following equation:

$$\begin{aligned}\text{window\_min}(t) &= \min\{p(t-N+1), p(t-N+2),..., p(t)\}\\ \text{window\_max}(t) &= \max\{p(t-N+1), p(t-N+2),..., p(t)\}\end{aligned} \quad (1)$$

$$\begin{aligned}\text{p\_min(t)} &= \min\{p\_\min(t-1), \text{window\_max}(t)\}\\ \text{p\_max(t)} &= \max\{p\_\max(t-1), \text{window\_min}(t)\}\end{aligned} \quad (2)$$

3) the median power $p(t)_m$ is computed as the median of the buffered frame energy values (*N*). This median power is used for comparison with a threshold.
4) the threshold is calculated as shown below:

$$\text{thr} = \text{p\_min} + k * (\text{p\_max} - \text{p\_min}) \text{, where } 0 < k < 1 \quad (3)$$

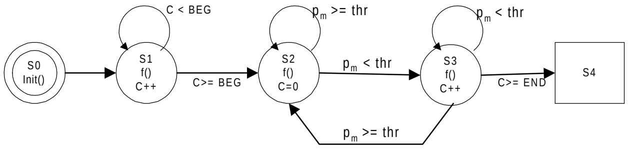

Figure 1 State machine that describes the rank order filtering scheme.

The algorithm to detect the end-of-utterance for each sub-band can be summarised as follows. First the counter is initialised to zero. The counter value is then increased for every frame in which $p(t)_m$ is less than an estimated threshold. If $p(t)_m$ stays long enough under the threshold, then it is assumed that speech has ended in that sub-band, and the end-of-utterance is detected in the current sub-band. If $p(t)_m$ exceeds the threshold, then the counter is reset to zero, and counting starts again (this means that speech is still detected in that sub-band). As shown in Figure 1, the comparison of $p(t)_m$ against the threshold is initiated after *'BEG'* number of frames. If the counter *C* exceeds *'END'*, then the recognition is stopped and state *S*4 is reached.

Figure 2 shows the plot of an energy contour of the lowest sub-band for the utterance "*Pekka Kapanen*" until the detection of the end-of-utterance. The figure also shows the contours of *p_min(t), p_max(t)* and the threshold.

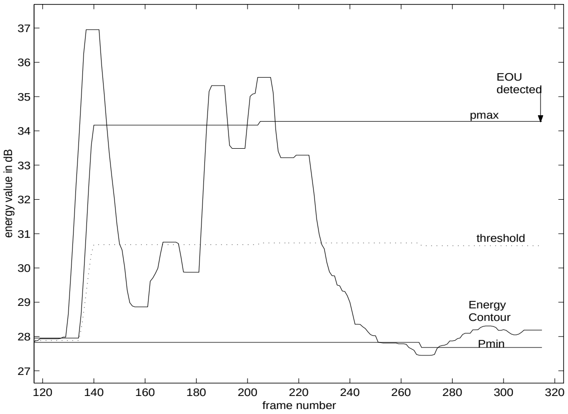

Figure 2 An example of the end-of-utterance detection in the lowest sub-band of the utterance "Pekka Kapanen".

3. EXPERIMENTS FOR ISOLATED WORD RECOGNITION

Experiments were initially carried out to test the end-of-utterance detection for isolated word recognition. The results were compared with a baseline confidence score based and full-band energy based EOU detection.

Some of the terminology used in the result tables is explained here. The target of the end-of-utterance detection system is to detect the time instant when speech ends with a sharp 0.80 seconds delay (80 frames in the test system). As shown in Fig. 3, ***early detections*** in this context mean that the end-of-utterance has been detected earlier than 400ms (40 frames) after the actual end of speech. Only when the end-of-utterance is detected before the actual speech ends, then some recognition errors due to end of utterance failure may occur. ***Late detection*** means that there is significantly longer delay than 0.8 seconds (greater than 1.2 seconds for isolated word detection and 1.35 seconds for continuous speech detection), which is not pleasant for the user, but does not necessarily deteriorate the recognition accuracy. The utterances for which the EOU algorithm fails to stop the recognition are termed as ***Algorithm failure***. Here the processing of the utterance continues until the end of the speech file is reached. It is to be noted here that all the utterance files used in the experiments had a large enough non-speech region at the end of the utterance. All the other utterances, which do not fall under any of the above three categories produced proper end-of-utterance detections and are classified as ***proper detections***. It should be noted that hand marked label files were used in detecting the actual end of speech for measurements.

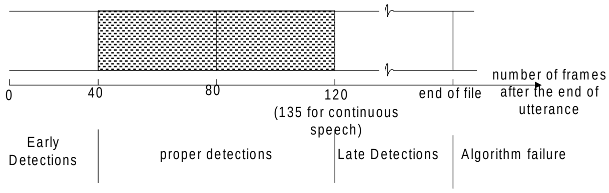

Figure 3 Explanation of terminology used in the results table.

3.1 Isolated Command Word Recognition

The isolated command word experiments were carried out for the following different schemes:

- Experiments using the confidence measure
- Experiments using single frequency band energy
- Experiments using multi-channel sub-band energy based end-of-utterance scheme

The database contained 60 confusable single and two-word Finnish isolated phrases, spoken by a single speaker with each utterance repeated 12 times. The tests were conducted by artificially adding both car noise (Volkswagen car driving at 115 km/h) and music (pop music by Bruce Springsteen) simultaneously to clean speech at different SNR combinations.

The first baseline experiments used the confidence measure for detecting the end-of-utterance. This confidence was based on the difference between the word model and the garbage model log-likelihood scores [3]. The confidence scores of each model was normalised with the length of the recognised word and compared against a pre-defined confidence threshold. The threshold was optimised to obtain the best results. Table 1 gives the results of the experiments with the optimised confidence threshold scheme. From the results, one can see that this scheme works well in clean and less noisy conditions, but fails poorly at low SNRs. The average percentage of utterances that do not get detected by the EOU algorithm is about 48%.

Noise (in dB)	Music (in dB)	Proper Detections (%)	Early Detections (%)	Late Detections (%)	Algorithm Failure (%)
-[1]	-	91.7	0.2	5.9	2.3
0	-	68.6	3.0	0.8	27.6
-5	-	23.5	2.4	0.2	73.9
-	10	65.9	0.9	2.1	31.1
0	10	31.5	1.5	0.5	66.5
-5	10	10.9	1.1	0.6	87.4
Average		48.7	1.5	1.7	48.1

Table 1 Results with the confidence based EOU scheme.

The results using the energy-based scheme with a single frequency band is given in Table 2, while Table 3 gives the results for the proposed multi-channel sub-band energy based EOU scheme. It can be seen from the results that the single channel energy based scheme provides better results than the confidence based method especially in very high car noise conditions, but performs badly in the presence of background music. If one takes a typical example of a person travelling in a car at 115 km/h with music in the background, represented by 0 dB car noise and 10 dB music, this scheme produces only 86.2% proper detections.

The best results are obtained with the multi-channel sub-band based scheme, which works well with both car noise and background music. There are very few *Algorithm failures* as compared to the previous schemes. In the example case (0 dB car noise and 10 dB music), the proposed algorithm provides 95.2% proper detections, a relative improvement of almost 65.2% over the single channel scheme. There is more than 93% overall proper detections using this multi-channel sub-band based end-of-utterance scheme.

Noise (in dB)	Music (in dB)	Proper Detections (%)	Early Detections (%)	Late Detections (%)	Algorithm Failure (%)
-	-	99.7	0.3	0.0	0.0
0	-	94.1	5.3	0.3	0.3
-5	-	87.6	6.8	1.8	3.8
-	10	88.2	0.6	3.9	7.3
0	10	86.2	5.0	3.3	5.5
-5	10	74.9	6.4	8.3	10.5
Average		88.4	4.1	3.0	4.5

Table 2 Results for single-channel energy based EOU scheme.

Noise (in dB)	Music (in dB)	Proper Detections (%)	Early Detections (%)	Late Detections (%)	Algorithm Failure (%)
-	-	100.0	0.0	0.0	0.0
0	-	97.4	2.4	0.2	0.0
-5	-	91.7	4.2	3.0	1.1
-	10	88.6	0.9	7.3	3.2
0	10	95.1	1.5	2.6	0.8
-5	10	87.3	2.3	7.0	3.5
Average		93.4	1.9	3.3	1.4

Table 3 Results using multi-channel sub-band energy based EOU scheme.

A comparison of the histograms of the time instants when the end-of-utterance was detected (given by the number of frames after the real end of the speech utterance), helps to illustrate the superior performance of the multi-channel based scheme over the conventional full band energy method. These histograms are shown here in Figure 4 for the particular case of speech mixed with 0 dB car noise and 10 dB music. Ideally, these methods should detect the end of an utterance at exactly 80 frames after the end of speech. It can be seen from the figure that there are too many early detections for the single channel scheme and also has a higher spread than the multi-channel method. On the other hand, the multi-channel end-of-utterance scheme is seen to have a very high concentration around the ideal 80 frames delay.

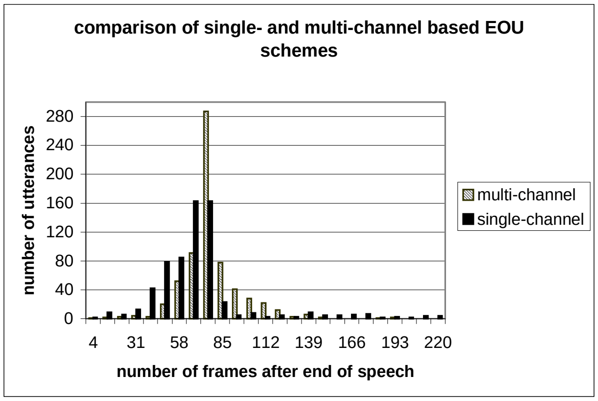

Figure 4 Histograms of the time instant after the end of speech when end-of-utterance is detected for single and multi-channel based schemes. This particular case involves speech mixed with 0 dB car noise and 10 dB music. The target is to detect the end-of-utterance after exactly 80 frames.

[1] '-' denotes that there is no noise addition

4. END-OF-UTTERANCE FOR CONTINUOUS SPEECH RECOGNITION

One of the major problems for detection of the end of continuous speech, as opposed to isolated speech, is the presence of inter-word pauses in the former. We have also observed that people often tend to decrease their speaking effort towards the end of a multi-word utterance in real-world conditions. Due to these, we expected some performance degradation when we decided to test the end-of-utterance scheme proposed in the previous section in the continuous speech recognition case. Indeed, the use of the isolated end-of-utterance algorithm produced more than 50% early detections, when applied to the recognition of spoken digit sequences.

The major change in the algorithm for continuous speech was the usage of the intermediate recognition results in addition to the individual sub-band energy information for detection of EOU. This recognition result gives the number of recognised words up to the given frame and gets updated as soon as a new word is recognised. Detection is disabled for a fixed number of frames after an intermediate result. This will reduce the risk of false early EOU detection. This requirement is combined with the earlier condition for isolated word EOU detection (*n* out of *M* sub-bands should trigger EOU) for stopping the recognition.

To cater to the changes in the speaking effort, shorter-term adaptation of the global maximum and minimum powers, used to compute the threshold, are carried out. The change in the buffered power values is used to determine the rate of adaptation for *p_max* and *p_min*. The maximum and minimum values of the buffered sub-band energy values (*win_max* and *win_min*) are computed, and compared with the current *p_max* and *p_min*. The updation of *p_max* and *p_min* are given by the following equations:

$$\begin{aligned}\text{p\_min} &= (1-\beta)\text{p\_min} + (\beta * \text{win\_min})\\ \text{p\_max} &= (1-\beta)\text{p\_max} + (\beta * \text{win\_max})\end{aligned} \qquad (4)$$

where $0 < \beta < 1$. The update coefficient β can be made bigger (or smaller) depending on whether the difference *p_min-win_min* or *win_max-p_max* is bigger (or smaller), to further improve the adaptation to varying speaking effort.

4.1 Speaker Independent Digit Dialler Experiments

The end-of-utterance for continuous speech recognition was tested for a speaker independent digit-dialling task. The test database consisted of Finnish connected digit utterances spoken in a clean environment. There were 375 speakers with an approximately equal number of males and females, with a total of 3686 six-digit utterances. During active listening, when the mobile phone waits for an activation phrase for the phone to be activated (to make a call), there is likely to be both stationary noise and music in the background [6]. However, when the phone is activated, people tend to reduce the volume of music during name or digit dialling. Hence, the digit dialling experiments, unlike the previous command word recognition tests were conducted only in stationary car noise environments.

The results of using the modified multi-channel sub-band scheme for end-of-utterance is shown in Table 4. The algorithm had been tuned to obtain very little early detections so that inter-word pauses are not detected as end of an utterance. It can be seen that the algorithm works well in both clean environment and in car noise conditions. The results in Table 4 show that on average, almost 94% of the connected digit utterances are detected properly by the proposed end-of-utterance scheme.

Noise Type	SNR (in dB)	Proper Detections (%)	Early Detections (%)	Late Detections (%)	Algorithm Failure (%)
-	-	97.6	1.3	0.4	0.7
Car	10	95.8	0.6	3.3	0.4
Car	5	94.8	0.9	3.9	0.4
Car	0	88.0	4.6	6.9	0.5
Average		94.0	1.8	3.6	0.5

Table 4 Results using the proposed sub-band EOU scheme for connected digit recognition task

5. CONCLUSIONS

We have proposed a robust multi-channel sub-band energy based end-of-utterance detector that works well in various realistic noise environments. The end-of-utterance detector is essential to improve the predictability and usability of a practical speech recognition system. It helps to provide a timely response to the user prompts. The proposed multi-channel method is shown to outperform the single band energy scheme with a relative average improvement in proper detection rate of more than 43% for a isolated command word recognition task. The scheme is found to be robust to both stationary and non-stationary noise conditions, which is essential for speech recognition applications like the use of hands-free voice activation in a car environment.

The algorithm was also adapted to work for detection of continuous speech. Experiments with a connected digit dialling task provided an average 94% proper end-of-utterance detections in a stationary car noise environment including very low SNRs.

REFERENCES

[1] S. Sagayama, K. Aikawa, "Issues Relating to the Future of ASR for Telecommunications Applications", *Proc. of ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 75-82, France,1997.

[2] L. Lamel, R. Rabiner, J. Rosenberg, J. Wilpon, “An improved end-point detector for isolated word recognition”, *IEEE ASSP Magazine*, pp. 777-785, 1981.

[3] O. Viikki, K. Laurila, P. Haavisto, "A Confidence Measure for Detecting Recognition Errors in Isolated Word Recognition", *Proc. International Conference on Speech Science and Technology*, pp. 67-72, Adelaide, Australia, 1996.

[4] H. Boulard, B. D'hoore, J.-M. Boite, “Optimizing recognition and rejection performance in wordspotting systems”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-373 - I-376, Australia 1994.

[5] M. G. Rahim, C.-H. Lee, B.-H. Juang, “Robust utterance verification for connected digits recognition”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 285- 288, Detroit, USA 1995.

[6] J. Iso-Sipilä, K. Laurila, R. Hariharan, O. Viikki, "Hands Free Voice Activation in Car Environments", *Proc. Eurospeech 97*, pp. 2375-2378, Rhodos, Greece, 1997.