Detection and Recognition of Number Sequences in Spoken Utterances

Guillermo Aradilla [a] Jitendra Ajmera [b]

IDIAP–RR 07-42

September 2007

published in
SiMPE 2007

[a] IDIAP Research Institute and Ecole Polytechnique Fédérale de Lausanne (EPFL)
[b] Deutsche Telekom Laboratories

IDIAP Research Report 07-42

Detection and Recognition of Number Sequences in Spoken Utterances

Guillermo Aradilla   Jitendra Ajmera

September 2007

published in
SiMPE 2007

Abstract. In this paper we investigate the detection and recognition of sequences of numbers in spoken utterances. This is done in two steps: first, the entire utterance is decoded assuming that only numbers were spoken. In the second step, non-number segments (garbage) are detected based on word confidence measures. We compare this approach to conventional garbage models. Also, a comparison of several phone posterior based confidence measures is presented in this paper. The work is evaluated in terms of detection task (hit rate and false alarms) and recognition task (word accuracy) within detected number sequences. The proposed method is tested on German continuous spoken utterances where target content (numbers) is only 20%.

1 Introduction

This paper presents a study on detection and recognition of a sequence of numbers in a spoken utterance. This task involves two steps as shown in Figure 1: a) detecting if there exists a sequence of numbers in the utterance and if so, b) recognition of the spoken numbers. It can be pointed out that the first task is similar to what is known in literature as keyword spotting (KWS) while the second task can be seen as a continuous speech recognition (CSR) problem. Therefore, techniques and evaluation methods employed in this work are influenced by both KWS and CSR approaches.

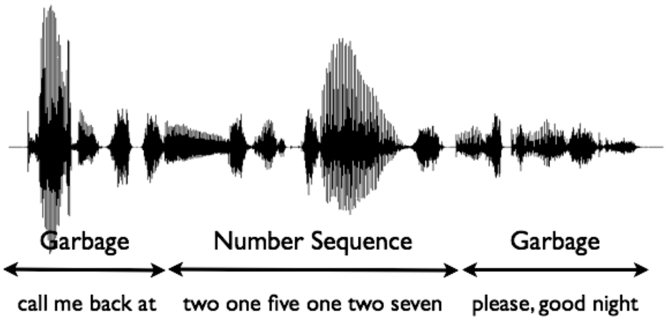


Figure 1: Given a spoken utterance, the task is to detect if there is a number sequence and if so, recognize the numbers that have been spoken.

This work can result in many useful applications. For example, detection of a phone number from a voice message which can be displayed immediately on a phone screen or voice activated dialing where users can say a phone number in the middle of a conversation. The algorithm explored in this work can be extended to detect other sequences of keywords such as street addresses, train names, etc.

In this kind of applications, one of the major challenges is to reliably reject non-target events and focus only on target segments of a given spoken utterance. This can be seen as what is referred in the KWS literature as "garbage modeling". Garbage models can be estimated from training data as presented in [5], where they use a loop of phone-based hidden Markov models (HMMs) to represent non-target events. A non-parametric approach is considered in [6], where the likelihood of the garbage model is computed in an online manner based on the average of the top N likelihoods from all phonemes. Also, approaches without explicit garbage models have been studied, e.g. a modified Viterbi algorithm is presented in [3].

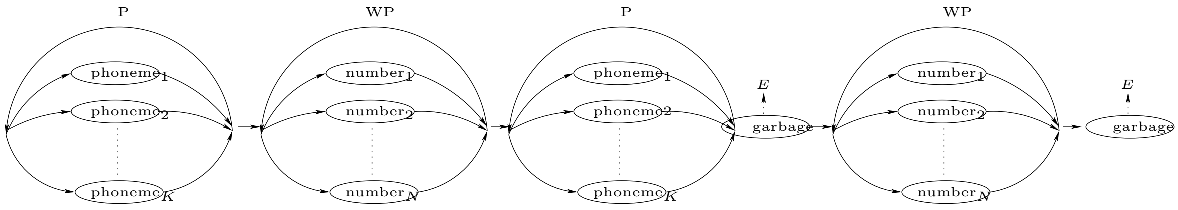
In this paper, we propose a new approach towards rejecting non-target events and compare it with conventional garbage modelling approaches. In this approach the whole utterance is first decoded using only the keyword lexicon and then non-keywords are discarded based on their confidence measures.

Different confidence measures have been previously studied which indicate reliability of a hypothesized word. For instance, duration normalized log-likelihoods is used for KWS in [9]. A method for utterance verification as post-processing based on log-likelihood ratio (LLR) is proposed in [4]. In this work, we analyze word confidence measures based on phoneme posterior probabilities [10, 7].

This paper is organized as follows: Section 2 describes the data that have been used for this work. Section 3 presents the different types of garbage models considered in this work. Confidence measures applied in our experiments are explained in Section 4. Finally, Section 5 concludes this paper.

2 Database

The German VeriDat database [8] has been used for this work. This database consists of continuous speech utterances spoken by 150 speakers in 20 different sessions, each session having 40 utterances.



(a) Loop of Phonemes (b) Normalized Garbage Model

Figure 2: Schemes of garbage models. It must be noted that both structures assume that a sequence of number exists and it is pronounced in the middle of the utterance. The word insertion penalty WP between numbers, penalty factor between phonemes P and constant value E have been tuned to maximize word recognition accuracy.

We have split the database into two categories: utterances containing numbers (number utterances) and the rest (garbage utterances). The number of words in number utterances varies from 4 to 16.

The set of keywords consists of 23 numbers pronounced in German. The keywords have been described by hybrid HMM/MLP (multi-layer perceptron) models [2] with uniform phoneme prior probabilities (scaled likelihoods have also been used with similar results). Considering that recognition has to be performed only on keywords (numbers in our case), acoustic model (MLP) has been trained using only number utterances. These training number utterances correspond to the first 100 speakers (25631 utterances).

The phonetic transcriptions of numbers correspond to a set of 28 phonemes, which is also the number of MLP outputs. A hidden layer of 1000 units has been used for the MLP. A context of 9 MFCC (mel frequency cepstral coefficient) feature vectors with delta and delta-delta features are used as input (39x9=351 input units) every 10ms.

The test utterances have been artificially created by concatenating garbage utterances belonging to the same speaker and session at the begin and end of number utterances. Only utterances from the last 50 speakers have been used for testing. Silence has been removed from the concatenated utterances to make them more natural.

As mention earlier, rejection of non-target events remains the most critical part of the task. In the next section, we compare different garbage models employed to achieve this.

3 Comparison of Garbage Models

In this section, description and comparison of the different garbage models used in our work are presented.

3.1 Description

- **Loop of Phonemes**. As described in [1], a loop of phoneme models can be used as garbage model. A scheme of this approach is shown in Figure 2(a). We use context-independent phoneme HMM which are also used to build the keyword models. When applying this model, a self-transition penalty P in the phoneme models is used to control the number of keyword detections and false alarms.

- **Normalized Garbage Model**. This model is similar to the approach described in [6]. The objective is to find a sequence of numbers $\{w_i\}_{i=1}^N$ such that their normalized score $NS(\{w_i\}_{i=1}^N)$ given by Equation 1 is higher than a threshold E.

$$NS(\{w_i\}_{i=1}^N) = \frac{1}{e-b+1} \sum_{t=b}^{e} \log p(q_t|x_t) \tag{1}$$

In Equation 1, b and e represent the beginning and ending frame of the hypothesized sequence of numbers and $p(q_t|x_t)$ is the posterior probability of the phoneme q_t given the acoustic frame x_t at time t. The phoneme sequence $\{q_t\}_{t=b_w}^{e_w}$ is given by the Viterbi path.

This criterion is equivalent to employing a garbage model whose emission probability is constant and equal to E, i.e. $P(G|x_t) = E$. This model is illustrated in Figure 2(b).

The constant value E can be considered as a tuning parameter to control the number of detected keywords and false alarms.

- **Proposed Method**. This approach does not use an explicit garbage model. The whole utterance is first decoded using the lexicon formed by the keywords. In this way, the entire utterance (including garbage parts) is hypothesized as a sequence of numbers. A second step is then required to discard words belonging to the garbage segments at begin and end of the utterance.

  A confidence measure is computed for each hypothesized word. Our method selects the longest sequence of numbers $\{w_i\}_{i=1}^{N}$ such that its average confidence measure is higher than T, i.e.

$$\frac{1}{N}\sum_{i=1}^{N} CM(w_i) > T \qquad (2)$$

  where $CM(w_i)$ is the confidence measure of word w_i. The detection of the sequence $\{w_i\}_{i=1}^{N}$ is done by following a greedy algorithm. T is a threshold that controls the number of detections and false alarms. A discussion of different confidence measures explored in this work is presented in Section 4.

3.2 Comparison

The sequence of numbers obtained as explained in the previous section is compared against the transcription of the "number only" part of the concatenated test utterances. The resulting word accuracy will have elements of false alarms and deletion errors as well as recognition errors within the number sequence.

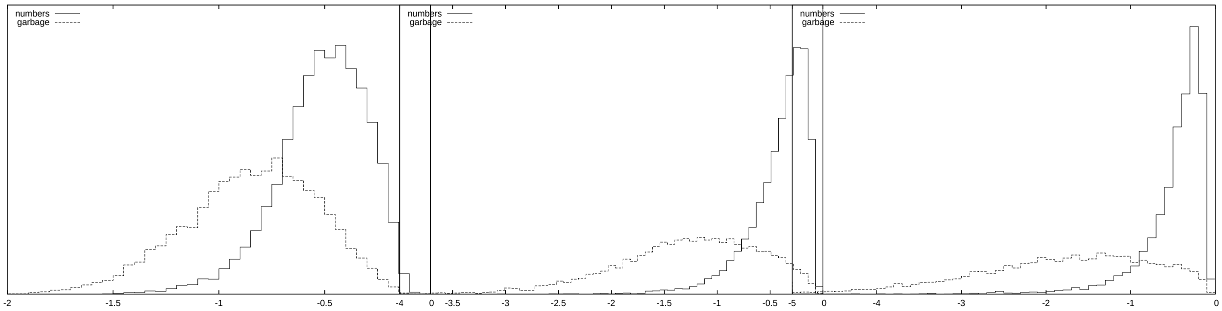
It should be noted that word accuracy corresponding to the CSR task applied on number utterances is 86.5%. This means that this is the maximum accuracy we can get for the current task on concatenated utterances. Another indicator of the task difficulty is that the percentage of target events (numbers) in the test concatenated utterances is only 40%.

The tuning parameters to control the number of detections and false alarms has been optimized for the individual approaches. Confidence measure ($CM(w_i)$ in Equation 2) used for this comparison is equal to $NS(w_i)$ in Equation 1, where b and e represent, in this case, the beginning and ending frame of word w_i. Table 1 presents the comparison of the three approaches in terms of word accuracy.

Loop of Phonemes	55.0%
Normalized Garbage Model	62.9%
Proposed Method	66.5%

Table 1: Word accuracy of the system using different garbage models.

Table 1 shows that the proposed approach performs better than the other garbage models. The proposed approach employs word confidence measures for discriminating between target (number) and non-target (garbage) events. Different confidence measures have been studied in this work for this purpose. Next section presents discussion and comparison of different confidence measures based on phoneme posterior probabilities.



(a) Entropy (b) Normalized posteriors (c) Phone normalized posteriors

Figure 3: Empirical distributions of the confidence measures. These distributions have been obtained from the histograms of the difference confidence measures from both number and garbage utterances.

4 Comparison of Confidence Measures

In this section, we describe and compare different word confidence measures to be used in the second step of the proposed method. These confidence measures are based on phoneme posterior probabilities obtained from a MLP as described in Section 2.

Posterior probabilities result in efficient confidence scores because of the inherent discriminative training of the MLP as well as the normalized nature of the MLP output [10]. In the following, we present confidence measures investigated for this task.

4.1 Description

- **Normalized Posteriors** (CM_{post}). This measure is defined as the average value of the logarithm of the posterior probabilities of the phonemes given by the Viterbi path. This is computed as:

$$CM_{post}(w) = \frac{1}{e_w - b_w + 1} \sum_{t=b_w}^{e_w} \log p(q_t|x_t) \tag{3}$$

  where b_w and e_w denote the beginning and ending frame of word w.

- **Phone Normalized Posteriors** (CM_{phone}). CM_{post} can be further normalized by the duration of the underlying phonemes to result in another confidence measure computed as following:

$$CM_{phone}(w) = \frac{1}{J_w} \sum_{j=1}^{J_w} \frac{1}{e_w^j - b_w^j + 1} \sum_{t=b_w^j}^{e_w^j} \log p(q_t|x_t) \tag{4}$$

  where b_w^j and e_w^j represent, respectively, the beginning and ending frame of the phoneme j and J_w denotes the number of phonemes forming the keyword w. This measure has been used in the previous section.

- **Entropy** (CM_H). As described in Section 2, the MLP to obtain the phone posterior probabilities has been trained only with utterances containing numbers. Hence, the MLP has only learnt the phonemes that constitute numbers. Therefore, the MLP can be expected to provide more confident output when the input belongs to a number than when it belongs to garbage. A measure to determine the confidence of the output generated by the MLP is entropy. Posterior probabilities estimated by the MLP can be seen as a discrete distribution over the phoneme space and entropy can thus be computed.

This confidence measure is defined as the minus average entropy value over keyword w.

$$CM_H(w) = \frac{1}{e_w - b_w + 1} \sum_{t=b_w}^{e_w} \sum_{k=1}^{K} p(q_k|x_t) \log p(q_k|x_t) \tag{5}$$

where K is the total number of phonemes.

The next subsection presents the comparison of these three confidence measures.

4.2 Comparison

The confidence measures are used to detect sequence of numbers in the decoded test concatenated utterances. As mentioned earlier, the proposed method detects the sequence with maximum number of words $\{w_i\}_{i=1}^N$ such that the average of their confidence measures is higher than a threshold T.

$$\frac{1}{N} \sum_{i=1}^{N} CM(w_i) > T \tag{6}$$

If there is no sequence of words $\{w_i\}_{i=1}^N$ that satisfies this condition, then the test utterance is considered not to contain a sequence of numbers. The threshold T controls the number of detections and false alarms at both the word level and utterance level. At the word level, these errors correspond to insertions and deletions within the "number only" part of the number utterances, whereas at utterance level, it provides accuracy of detecting utterances containing numbers (hit and false alarm rates).

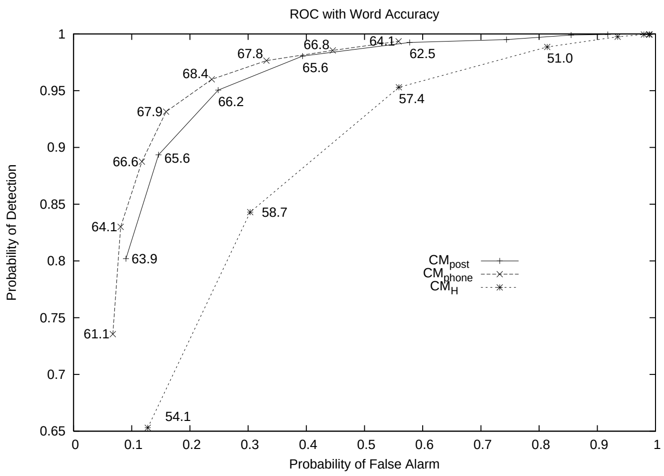


Figure 4: ROC curves for different confidence measures. The threshold T controls the probability of detections and corresponding word accuracy. The length of the detected number sequence is constrained to have at least 4 words.

Moreover, a constraint of a minimum number of words in the detected sequence can be easily imposed. If the longest detected sequence $\{w_i\}_{i=1}^N$ does not contain a specified minimum number of words, then that utterance is discarded. Since number sequences in the database contain at least 4 words, the constraint of a minimum number of 4 words has been applied in the comparison of this section. In the next section, we investigate the effect of the length of number sequences on the performance of the system.

We present a comparison of different confidence measures using both receiver operating characteristic (ROC) curve and word accuracy criteria. For this comparison, 2000 utterances containing a sequence of numbers in the middle (concatenated utterances) and 2000 garbage utterances have been used. In these 4000 test utterances, the percentage of target events is only 20% and this provides an indication of the difficulty of task.

Figure 3 presents the distributions of the three confidence measures obtained from number and garbage utterances in training data. The overlapping area between the two distributions is an indicator of the lack of discrimination of the confidence measure. From the figure, we can observe that CM_H presents the highest overlapping, whereas CM_{phone} is slightly more discriminant than CM_{phone}.

Figure 4 shows a comparison in terms of ROC curve and word accuracy controlled by the threshold T. Word accuracy has been computed on the set of utterances that have been correctly classified. The following observations can be drawn from these results:

- CM_{phone} is the confidence measure that presents the best results both in terms of ROC curve and word accuracy. This is coherent with observations made in [7].
- The better performance of CM_{phone} when compared to CM_{norm} can be explained by the fact that long phonemes contribute more to CM_{norm} than short phonemes. This effect may yield unreliable results when there is a mismatch between acoustic features and the keyword model. In this case, incorrect phonemes are assigned a short duration and contribute little to the confidence measure. However, in CM_{phone} every phoneme contributes with the same weight independently of its duration.
- CM_H appears not to be a very reliable measure mainly because the phonemes appearing in the number transcriptions correspond to more than half the total number of German phonemes. Hence, entropy is still low for most of the phonemes appearing in garbage words.
- A better probability of detection and false alarm corresponds to a better word accuracy. This implies that recognition and detection are two inter-related tasks.
- For each confidence measure, a value of T that results in a better compromise between detection and false alarms also results in a better recognition accuracy.

4.3 Experiments with Sequences of Different Lengths

It can be expected that the longer the sequence of numbers, the easier should be the detection of utterances containing number sequences. Also, as mentioned in the previous section, recognition of these number sequences should also result in higher word accuracies.

In this section, we impose another constraint while detecting the sequence of numbers. Accordingly, only sequences containing a minimum number of words (numbers) need to be detected by the system. Thus, the minimum number of words can be seen as a hyper-parameter (or input) to the system.

CM_{phone} is used as word confidence measure for these experiments. In the second step of the proposed method, if the longest detected sequence of numbers does not contain the specified minimum number of words, then that utterances is discarded. Experimental results with 4, 7, 11 and 16 as minimum number of words are presented in Figure 5. As expected, both ROC curve and word accuracy improve when increasing the minimum length of the sequences to be detected.

5 Conclusion

In this paper we have investigated detection and recognition of sequences of numbers in spoken utterances. This task and its evaluation involve approaches corresponding to both keyword spotting and continuous speech recognition problems.

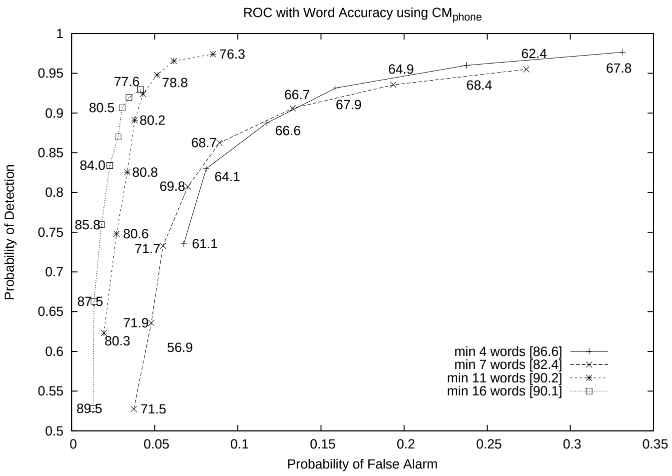


Figure 5: ROC curves and word accuracy for CM_{phone} using the constraint of a minimum number of words. The number between brackets in the legend represents the best accuracy that can be obtained. This accuracy has been obtained by applying normal Viterbi decoding on the number utterances.

A method was proposed to reliably reject non-number segments (garbage) and find optimal sequence in the utterance. The proposed method was compared against other conventional garbage modeling approaches and found to be better. The proposed approach consists of first decoding entire utterance assuming that there were only numbers spoken. The second step consists of finding the confidence measure for each hypothesized word and then discarding those words with lower confidence scores.

Several phoneme posterior probability based word confidence measures were compared for this purpose. Phone duration normalized confidence measure was found to be the best.

It was observed that detection and recognition are interrelated tasks and better performance on detection task generally results in better recognition accuracies and vice versa.

Furthermore, study of the system performance as function of minimum number of words to detect was also carried out. Word accuracies of 68.4, 71.9, 80.8 and 89.5 percent were observed corresponding to minimum number of words being 4, 7, 11, 16, respectively.

6 Acknowledgements

This work was supported by the joint research project number 2006/DT-1 between IDIAP Research Institute and Deutsche Telekom AG Laboratories. The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on “Interactive Multimodal Information Management (IM2)”

The authors also would like to thank Mathew Magimai Doss and Hervé Bourlard for useful discussions during this work.

References

[1] H. Bourlard, B. D’hoore, and J.-M. Boite. Optimizing recognition and rejection performance in wordspotting systems. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, 1994.

[2] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*, volume 247. Kluwer Academic Publishers, Boston, 1993.

[3] J. Junkawitsch, L. Neubauer, H. Höge, and G. Ruske. A new keyword spotting algorithm with pre-calculated optimal thresholds. *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 4:2067–2070, 1996.

[4] E. Lleida and R. C. Rose. Utterance verification in continuous speech recognition: Decoding and training procedures. *IEEE Transactions on Speech and Audio Processing*, 8(2):126–139, 2000.

[5] R. Rose and D. Paul. A hidden markov model based keyword recognition system. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 129–132, 1990.

[6] M. C. Silaghi and H. Bourlard. Iterative Posterior-Based Keyword Spottin without Filler Models. *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 1999.

[7] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems," *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1990.

[8] U. Turk and F. Schiel. Speaker verification based on the german veridat database. In *Eurospeech*, pages 3025–3028, 2003.

[9] M. Weintraub. Improved keyword-spotting using sri's decipher large-vocabulary speech-recognition system. *Proceedigs of Human Language Technology*, 1993.

[10] G. William and S. Renals. Confidence Measures from Local Posterior Estimate. *Computer, Speech and Language*, 13:395–411, 1999.