THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Handling overlaps in spoken term detection

Citation for published version:
Wang, D, Evans, N, Troncy, R & King, S 2011, Handling overlaps in spoken term detection. in Proc. International Conference on Acoustics, Speech and Signal Processing. pp. 5656-5659. https://doi.org/10.1109/ICASSP.2011.5947643

Digital Object Identifier (DOI):
10.1109/ICASSP.2011.5947643

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proc. International Conference on Acoustics, Speech and Signal Processing

Publisher Rights Statement:
Wang, D., Evans, N., Troncy, R., & King, S. (2011). Handling overlaps in spoken term detection. In Proc. International Conference on Acoustics, Speech and Signal Processing. (pp. 5656-5659). doi: 10.1109/ICASSP.2011.5947643

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 13. Nov. 2019

HANDLING OVERLAPS IN SPOKEN TERM DETECTION

Dong Wang, Nicholas Evans, Raphaël Troncy

EURECOM
Multimedia Department
BP 193, F-06904
Sophia Antipolis, France

Simon King

Centre for Speech Technology Research
University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB, UK

ABSTRACT

Spoken term detection (STD) systems usually arrive at many overlapping detections which are often addressed with some pragmatic approaches, e.g. choosing the best detection to represent all the overlaps. In this paper we present a theoretical study based on a concept of acceptance space. In particular, we present two confidence estimation approaches based on Bayesian and evidence perspectives respectively. Analysis shows that both approaches possess respective advantages and shortcomings, and that their combination has the potential to provide an improved confidence estimation. Experiments conducted on meeting data confirm our analysis and show considerable performance improvement with the combined approach, in particular for out-of-vocabulary spoken term detection with stochastic pronunciation modeling.

Index Terms— Confidence measurement, stochastic pronunciation modeling, spoken term detection, speech recognition

1. INTRODUCTION

Spoken term detection (STD), as defined by NIST in 2006 [1], aims to provide for the searching of large quantities of audio without the need for reprocessing the audio signal every time a query is performed. The evaluations organized by NIST have attracted broad interest, including [2, 3, 4, 5, 6]. A typical STD system consists of an automatic speech recognition (ASR) component to transcribe speech signals into word or subword lattices and a detection component to search for occurrences of search terms within the generated lattices.

A well known problem of STD is the ubiquitous overlap among resulting detections. In most STD systems, overlapping detections are merged into a single detection by some *pragmatic* approach, for example, choosing the best detection as the merged detection, or accumulating the time and/or confidence measures of all the overlaps as the time and confidence measure of the merged detection [4, 6]. Frame-based treatments have also been reported [4], though they tend to have weak theoretical foundation.

Another challenge to existing overlap treatment stems from out-of-vocabulary (OOV) terms. We have shown in previous studies [7] that an ATWV-oriented decision strategy [8] is essential for OOV STD, however accumulated confidences are not normalized and thus cannot be applied together with ATWV-oriented decisions, although normalization techniques may provide some compensation [7]. Furthermore, stochastic pronunciation modeling (SPM) [9], which has been shown to be highly effective for OOV term detection, may further complicate the pattern of overlaps since more pronunciation candidates are taken into account.

In this paper, we present a theoretical reasoning for overlap treatment based on a concept of acceptance space. Specifically, we present an average time estimation and two confidence estimation approaches based on the Bayesian perspective and the evidence perspective respectively. The two confidence estimation approaches are then combined to give a normalized confidence measurement which is consistent with the ATWV-oriented decision. In the next section, we first present a statistical analysis of the pattern of overlaps for in-vocabulary (INV) terms and OOV terms, and then present time and confidence estimation approaches in Section 3. Experiments and results are reported in Section 4.

2. OVERLAP STATISTICS

This section reports a study of the different patterns of overlapping detections for INV and OOV terms. We first introduce the data used in the experiments and the reference system used to conduct STD.

2.1. Data and reference system

OOV terms are strictly defined as those terms which do not contain words in the system dictionaries nor in the training material for either the acoustic models (AM) or language models (LM). We selected 412 OOV terms from the AMI dictionary that do not occur in the COMLEX dictionary (published by LDC in 1996 and therefore historical from an STD perspective), and added another 70 *artificial* OOV terms (which occur more frequently) that are plausible search terms. This results in 482 search terms which have a total of 2736 occurrences in the evaluation data. These terms were removed from the system dictionaries and the speech and text training corpora. In addition, 256 INV terms which are mostly person and city names were chosen to perform a comparative study.

The speech data used in this work are from multi-participant meetings recorded using individual head-mounted microphones. After OOV purging, 122744 utterances (80.2 hours) of speech is available to train the AM. The NIST RT04s development set was used for parameter tuning. The evaluation set comprised the RT04s and RT05s evaluation sets and a meeting corpus recorded recently at the University

	# clusters	overlap ratio	max overlap
INV	32,436	235	177,608
OOV/1-best	75,854	82	145,440
OOV/SPM	767,714	73	158,400

Table 1. Statistics of overlap clusters.

	# clusters	overlap ratio	max overlap
INV	32,436	66	7,225
OOV/1-best	75,854	41	3,841
OOV/SPM	767,714	65	13,838

Table 2. Statistics of overlap clusters after merging strict overlaps.

of Edinburgh in the AMIDA project, totalling 11 hours of speech. The text corpus used to train the language model was provided by the AMI project and is the same as that used by the AMI RT05s large vocabulary continuous speech recognition (LVCSR) system [10]. It involves 521.4 million words in total after OOV purging. A 50k word dictionary from the AMI project (also OOV purged) was used to convert the word-based text corpus to a phoneme-based one. The same dictionary was used to train a joint-multigram model which is used for predicting pronunciations of OOV terms.

We built a phoneme-based STD system using the resources described above. The acoustic models are 3-state triphone HMMs employing conventional 39-dim MFCC features, with cepstral mean and variance normalisation (CMN + CVN) applied. A 6-gram phoneme LM was used to perform speech decoding (the LM order was optimized empirically). The averaged density of the resulting lattices is 805 nodes per second. The HTK toolkit was used to train the acoustic models and transcribe speech to lattices, and the SRI LM toolkit was used to train the phoneme 6-gram model. The term detector was implemented with *Lattice2Multigram* generously provided by the Speech Processing Group, FIT, Brno University of Technology [4]. Confidence normalization [7] was applied in all experiments.

2.2. Statistical analysis

To analyze the pattern of overlaps, we conducted STD for the INV terms and OOV terms respectively. We tested two scenarios for OOV terms: one with 1-best pronunciations and the other with multiple pronunciations based on SPM [9]. We define a maximum group of overlapping detections as a *cluster*, the averaged number of detection per cluster as the 'overlap ratio', and the size of the largest cluster as 'max overlap'. The statistics are presented in Table 1. It is clear that INV terms have a larger overlap ratio than OOV terms, indicating a greater overlap tendency for INV terms than for OOV terms. This can be explained by the fact that INV terms are better represented by the language model than OOV terms which leads to more dense INV paths than OOV paths in lattices. In addition, we find that more detections/clusters are obtained with SPM and the overlap ratio is slightly reduced, while the max overlap is slightly higher compared to the 1-best detection approach. This indicates that some clusters are enriched by detections based on alternative pronunciations, while the new added clusters tend to be small. Detailed investigation on the clusters generated by SPM confirms this conjecture.

Also of interest is how many of these overlaps are 'strict overlaps'. For strict overlaps we refer to detections with the same starting and ending time. Strict overlaps can be easily merged into a single detection through confidence accumulation, as we discuss in the next section. Table 2 shows the statistics after all strict overlaps are merged. Interestingly the overlap ratio of INV terms is significantly decreased, while this is not the case for OOV terms, especially when SPM is applied. This indicates that most of the overlaps for INV terms are strict overlaps. Again, it can be explained by the denser paths of INV terms than OOV terms.

3. OVERLAP TREATMENT

3.1. Acceptance space

The confidence of a detection is usually formulated as a detection posterior probability, given by:

$$c \quad = \quad P(K_{t_s}^{t_e}|O) \tag{1}$$

where $K_{t_s}^{t_e}$ denotes the event that search term K appears between time t_s and time t_e in the audio stream O. This confidence measure is usually computed from the lattices, and thus is referred to as the *lattice-based confidence*. The same formulation is adapted for SPM, although a hidden variable is introduced to represent possible pronunciations.

Scrutinizing (1), we see that it actually represents the confidence measure that a term K appears in a *specific* speech segment, i.e., a segment whose starting and ending times are precisely specified. However, detections hypothesized by an STD system are never precise, and factors such as imperfect acoustic modeling, ambient noise and limited vocabulary always impose uncertainty which leads to biased time segmentation. In fact, even manually labeled transcripts are not absolutely accurate. To address the inaccuracy in time segmentation, an *endurance level* is typically applied when evaluating STD performance. For example, in NIST evaluations, the endurance level is set to 0.5 seconds from the mid-point of a detection to the time span of the true occurrence; in the HTK tool HResults, the endurance level is set such that the starting and ending time of a hit detection should be located before and after the mid-point of the true occurrence respectively.

No matter how it is defined, the endurance level actually forms a small 'acceptance space'. Detections falling in this space are considered as hits, while ones outside this space are considered to be false alarms. The task of STD therefore amounts to searching for as many putative detections as possible in the acceptance space, and the decision process is the task of inferring the true occurrence. In real applications, since the acceptance space of an occurrence is unknown, we can assume a cluster of overlapping detections form the acceptance space of a potential occurrence and estimate its time and confidence measure by the detections in that cluster.

3.2. Time estimation

Assuming a cluster presents a possible occurrence, and that the overlapping detections are samples of this occurrence, then the time span of all the overlapping detections can be

regarded as an approximation of the entire acceptance space of the hypothesized occurrence. Using this approximation to estimate the time segment of the hypothesized occurrence is the so called *group time* approach. A better time estimation is achieved by assuming that the confidence measure of a detection is also the confidence measure that its time segment estimates that of the hypothesized occurrence, which leads to an *average time* approach formulated as follows:

$$s = \frac{\sum_i c(i)s(i)}{\sum_i c(i)} \tag{2}$$

where s is the estimated time segment, and where $s(i)$ and $c(i)$ are the time segment and confidence measure of the i-th detection respectively. If one particular detection is dominant (i.e. its confidence score is much higher than that of the others), then the average time reduces to the *best time* approach, which selects the time of the detection with the highest confidence measure as the time of the hypothesized occurrence.

3.3. Confidence estimation

We propose two confidence estimation approaches based on a Bayesian perspective and an evidence perspective respectively, and then combine them for a normalized confidence measurement that is consistent with ATWV-oriented decision making and is therefore more suitable for OOV STD.

Bayesian approach

Assuming that the detections of an overlap cluster are samples of a single occurrence and that their starting time t_s and ending time t_e are random, we derive the confidence of a term K appearing in a time span τ as follows:

$$\begin{aligned} c &= P(K,\tau|O) && (3)\\ &= \sum_{t_s,t_e} P(K,t_s,t_e|O) && (4)\\ &= \sum_{t_s,t_e} P(K|O,t_s,t_e)P(t_s,t_e|O) && (5) \end{aligned}$$

where $P(K|O,t_s,t_e)$ is the lattice-based confidence in (1), and $P(t_s,t_e|O)$ can be regarded as the prior probability that the time span (t_s,t_e) falls in the acceptance space of the hypothesized occurrence. Any form of distribution $P(t_s,t_e|O)$ can be assumed; if we assume a uniform prior, then the commonly used *accumulated confidence* is derived:

$$c_{acc} = \sum_{t_s,t_e} P(K|O,t_s,t_e) \tag{6}$$

Furthermore, assuming a single dominant detection in the cluster, we obtain the *best confidence* approach:

$$c_{best} = max_{t_s,t_e} P(K|O,t_s,t_e) \tag{7}$$

Although the above equations are derived from single pronunciations, it is straightforward to extend them for multiple pronunciations (as in the case of SPM) by treating the pronunciation as an additional hidden variable.

Evidence approach

Now let us assume that each detection provides a piece of 'evidence' for the hypothesized occurrence, and that all detections are independent events. Furthermore assuming that an occurrence can be ascertained if any of the detection events appears, we derive the following *exclusive evidence*:

$$c_{env} = 1 - \prod_i (1 - c(i)). \tag{8}$$

Combined approach

The Bayesian approach has a solid theoretical foundation and imposes no artificial assumptions, however it is difficult to specify the prior distribution. Even with a uniform distribution, estimating the normalization factor $P(t_s,t_e|O)$ remains a problem. Unnormalized confidence measures, such as the accumulated confidence, are not consistent with ATWV-oriented decision making and might be difficult to compensate with confidence normalization, which is particularly problematic for OOV STD. The evidence approach, on the other hand, results in normalized confidences, but the independence assumption is perhaps too strong.

A better solution is to combine these two approaches: first merge all strict overlaps by the Bayesian approach, and then merge the non-strict overlaps by the evidence approach. The idea is to marginalize out pronunciations and phone segmentations in the first step by keeping segment ending points fixed, and then estimate the confidence by assuming that the non-strict overlaps are independent. The resulting confidence is normalized, and is referred to as the *exclusive accumulated confidence*, given by

$$c_{eacc} = 1 - \prod_{t_1,t_2} (1 - \sum_{t_s^i=t_1, t_e^i=t_2} c(i)) \tag{9}$$

where t_s^i and t_e^i are the starting and ending time of the i-th detection respectively.

4. EXPERIMENTAL RESULTS

We present our experimental results with the proposed overlap treatment. The performance of three time estimation approaches (best time, group time and average time) together with four confidence estimation approaches (best confidence c_{best}, accumulated confidence c_{acc}, exclusive evidence c_{env} and exclusive accumulated confidence c_{eacc}) is reported in terms of average term weighted values (ATWV) [1].

4.1. Single pronunciation systems

Results for INV terms and OOV terms with 1-best pronunciations are reported in Table 3 and 4 respectively. The first observation is that neither the accumulated confidence nor the exclusive evidence shows any advantage over the simple best confidence. This is especially true for the exclusive evidence, for which results are poor in the case of INV STD, indicating that the dominant strict overlaps are far from independent. Second, the exclusive accumulated confidence provides slightly better performance than the best confidence. Comparing different time estimation approaches, the average time

shows a small advantage for INV terms but performs almost the same as the best time for the OOV terms, suggesting that a large number of non-strict overlaps are required for the average time estimation. For both INV and OOV terms, the group time performs consistently worse than the other two approaches. All these results support the theoretical analysis presented in the previous section.

	ATWV		
	Best time	Group time	Average time
c_{best}	0.5320	0.5312	0.5324
c_{acc}	0.5302	0.5293	0.5305
c_{env}	0.4951	0.4942	0.4954
c_{eacc}	0.5356	0.5347	**0.5359**

Table 3. STD results with various overlap treatment approaches for INV terms.

	ATWV		
	Best time	Group time	Average time
c_{best}	0.2911	0.2902	0.2910
c_{acc}	0.2910	0.2903	0.2910
c_{env}	0.2900	0.2891	0.2899
c_{eacc}	**0.2931**	0.2924	0.2931

Table 4. STD results with various overlap treatment approaches for OOV terms with 1-best pronunciations.

4.2. SPM system

Now we extend the 1-best OOV detection to SPM. Note that the accumulation step in the exclusive accumulated confidence estimation is a slightly more complex: strict overlaps based on various pronunciations should all be merged. The ATWV results are shown in Table 5. We see that both the accumulated confidence and exclusive evidence perform considerably better than the best confidence. This seems to suggest that the overlapping detections introduced by SPM tend to provide valuable information. As expected, the exclusive accumulated confidence provides the best performance. The average time and the best time show similar performance, and the group time performs the worst.

	ATWV		
	Best time	Group time	Average time
c_{best}	0.3479	0.3462	0.3478
c_{acc}	0.3552	0.3537	0.3552
c_{env}	0.3503	0.3485	0.3502
c_{eacc}	**0.3605**	0.3588	0.3604

Table 5. STD results with various overlap treatment approaches for OOV terms with SPM.

5. CONCLUSIONS

This paper presents a theoretical analysis for overlap treatment in STD. An average time estimation and two confidence estimation approaches are proposed based on the idea of an acceptance space. The theoretical reasoning and experimental results show that the best time is good enough for time estimation, and that the exclusive accumulated confidence is the best approach for confidence estimation. Note that this conclusion is slightly different if the evaluation is based on FOM values computed with HTK, in which case the group time always performs the best, due to the specific criterion of HTK for asserting hits.

6. ACKNOWLEDGMENTS

This work was partially supported by the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966, Collaborative Annotation for Video Accessibility (ACAV) and by the Adaptable Ambient Living Assistant (ALIAS) project funded through the joint national Ambient Assisted Living (AAL) programme.

7. REFERENCES

[1] NIST, *The spoken term detection (STD) 2006 evaluation plan*, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edition, September 2006.

[2] Murat Akbacak, Dimitra Vergyri, and Andreas Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition systems," in *Proc. ICASSP'08*, Las Vegas, Nevada, USA, March 2008, pp. 5240–5243.

[3] Jonathan Mamou and Bhuvana Ramabhadran, "Phonetic query expansion for spoken document retrieval," in *Proc. Interspeech'08*, Brisbane, Australia, September 2008, pp. 2106–2109.

[4] Igor Szöke, Michal Fapšo, Lukáš Burget, and Jan Černocký, "Hybrid word-subword decoding for spoken term detection," in *Proc. Speech search workshop at SIGIR (SSCS'08)*, Singapore, 2008, Association for Computing Machinery.

[5] Dimitra Vergyri, Izhak Shafran, Andreas Stolcke, Ramana R. Gadde, Murat Akbacak, Brian Roark, and Wen Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 2393–2396.

[6] Dogan Can, Erica Cooper, Abhinav Sethy, Chris White, Bhuvana Ramabhadran, and Murat Saraclar, "Effect of pronunciations on OOV queries in spoken term detection," in *Proc. ICASSP'09*, Taipei, Taiwan, April 2009, pp. 3957–3960.

[7] Dong Wang, Simon King, Joe Frankel, and Peter Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Proc. Interspeech'09*, Brighton, UK, September 2009, pp. 2139–2142.

[8] David R. H. Miller, Michael Kleber, Chia lin Kao, Owen Kimball, Thomas Colthurst, Stephen A. Lowe, Richard M. Schwartz, and Herbert Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 314–317.

[9] Dong Wang, Simon King, and Joe Frankel, "Stochastic pronunciation modeling for out-of-vocabulary spoken term detection," *IEEE Trans. on Audio, Speech, and Language Processing*, 2010.

[10] Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, Mike Lincoln, Jithendra Vepa, and Vincent Wan, "The AMI meeting transcription system: Progress and performance," in *Machine Learning for Multimodal Interaction*, vol. 4299/2006, pp. 419–431. Springer Berlin/Heidelberg, 2006.