Direct Posterior Confidence for Out-of-Vocabulary Spoken Term Detection

Dong Wang, Nuance Communications
Simon King, Joe Frankel, University of Edinburgh
Ravichander Vipperla, Nicholas Evans, Raphaël Troncy, EURECOM

Spoken term detection (STD) is a key technology for spoken information retrieval. As compared to the conventional speech transcription and keyword spotting, STD is an open-vocabulary task and has to address out-of-vocabulary (OOV) terms. Approaches based on subword units, e.g. phones, are widely used to solve the OOV issue; however, performance on OOV terms is still substantially inferior to that of in-vocabulary (INV) terms. The performance degradation on OOV terms can be attributed to a multitude of factors. One particular factor we address in this paper is the unreliable confidence estimation caused by weak acoustic and language modeling due to the absence of OOV terms in the training corpora. We propose a direct posterior confidence derived from a discriminative model, such as a multi-layer perceptron (MLP). The new confidence considers a wide-range acoustic context which is usually important for speech recognition and retrieval; moreover, it localizes on detected speech segments and therefore avoids the impact of long-span word context which is usually unreliable for OOV term detection.

In this paper we first develop an extensive discussion about the modeling weakness problem associated with OOV terms, and then propose our approach to address this problem based on direct poster confidence. Our experiments carried out on spontaneous and conversational multi-party meeting speech, demonstrate that the proposed technique provides a significant improvement in STD performance as compared to the conventional lattice-based confidence, in particular for OOV terms. Furthermore, the new confidence estimation approach is fused with other advanced techniques for OOV treatment, such as stochastic pronunciation modeling and discriminative confidence normalization. This leads to an integrated solution for OOV term detection that results in a large performance improvement.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: speech recognition, spontaneous speech search, spoken term detection

ACM Reference Format:
Wang, D., King, S., Frankel, J., Vipperla R., Evans N., Troncy, R. , 2011. Direct Posterior Confidence for

---

This work was carried out while Dong Wang was a Fellow on the EdSST interdisciplinary Marie Curie training programme at CSTR, University of Edinburgh and was extended while he was in EURECOM. The revision was taken when he was in Nuance. This work used the Edinburgh Compute and Data Facility which is partially supported by eDIKT, and has been 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.
Author's addresses: D. Wang, Nuance Communications, Kackert Street 10, Aachen, Germany; S. King and J. Frankel, Centre for Speech and Technology Research (CSTR), University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK; R.Vipperla and N. Evans and R. Troncy, Multimedia Department, EURECOM, BP 193, F-06904, Sophia Antipolis, France.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

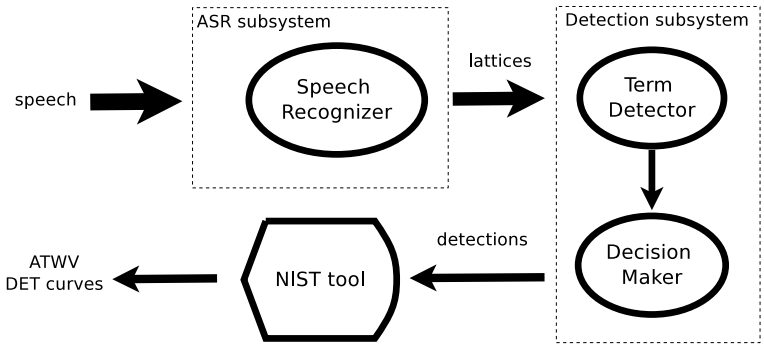

Fig. 1. The standard STD architecture: a speech recognizer converts speech into word/subword lattices; a term detector searches these lattices for potential occurrences of the search terms; a decision maker decides whether a detection is reliable. The NIST tool is used to evaluate detection performance.

Out-of-Vocabulary Spoken Term Detection. ACM Trans. Inf. Syst. 1, 1, Article 0 (January 2011), 34 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The ever-increasing volume of speech material on the web creates the need for spoken information retrieval (IR) techniques. Spoken term detection (STD), as defined by the National Institute of Standards and Technology (NIST), is a key technology for information retrieval from spoken documents, e.g., [Jones et al. 1996; Itoh et al. 2006; Olsson and Oard 2009]. According to NIST, STD aims to facilitate searching of vast and heterogeneous audio archives for occurrences of spoken terms without the need for reprocessing the audio signal for each query [NIST 2006]. The evaluation series organized by NIST has generated wide interest and attracted participation from several research groups including [Vergyri et al. 2007; Akbacak et al. 2008; Mamou and Ramabhadran 2008; Szöke et al. 2008; Can et al. 2009; Wang 2009; Natori et al. 2010; Lee and Lee 2010; Wallace et al. 2010; Jansen et al. 2010; Motlicek et al. 2010; Kaneko and Akiba 2010; Chan and Lee 2010; Meng et al. 2010; Schneider et al. 2010; Parada et al. 2010; Tejedor et al. 2010].

A typical STD system, as illustrated in Fig. 1, consists of two components. First, an automatic speech recognition (ASR) subsystem transcribes speech signals into intermediate representations, usually word or subword lattices, followed by a detection subsystem that searches for occurrences of the search terms. The latter comprises a term detector that searches the generated lattices for all potential occurrences of a search term, and a decision making component that determines if a potential occurrence is reliable enough to be hypothesized as a detection output. It is important to note that the ASR subsystem is run only once on the audio archives and that the detection subsystem has access only to the lattices and not the original audio.

In STD, a hypothesized occurrence is referred to as a *detection*; if the detection corresponds to an actual occurrence, it is called a *hit*, otherwise it is a *false alarm (FA)*. If an actual occurrence is not detected, it is referred to as a *miss*. We define a detection of a search term K as *a finding of a partial path in the lattice that represents* K, and denote it as a tuple d that encapsulates all the information available for this detection:

$$d = (K, \tau = (t_s, t_e), v_a, v_l, ...) \tag{1}$$

where v_a, v_l represent the acoustic and language model scores respectively, and τ denotes the time boundaries of the speech segment (from time t_s to t_e) where the de-

tection resides. Other informative factors such as pronunciation probability that we present shortly are denoted by "...".

Each putative detection is assigned a confidence score, or simply a *confidence*. Accurate confidence estimation is not only important for determining reliable putative detections in STD, but also for deriving relevance decision in spoken IR, e.g., [Jones et al. 1996; Olsson and Oard 2009]. Letting $K_{t_s}^{t_e}$ denote the event that the term K appears in the speech segment starting at time t_s and ending at time t_e, the confidence of $d = (K, \tau = (t_s, t_e), ..)$ can be evaluated by the posterior probability that the event $K_{t_s}^{t_e}$ appears given speech O. This is formulated as:

$$c(d) = P(K_{t_s}^{t_e}|O) \tag{2}$$

where $c(d)$ denotes the confidence of the detection d. A popular implementation of confidence estimation is based on lattice posterior probabilities which are derived from acoustic and language model scores via the Bayes rule, as will be discussed in detail in Section 3.1.

A particular feature that discriminates STD from other ASR-based tasks such as speech transcription and keyword spotting is that an open vocabulary is assumed for the former. Queries may thus contain words that are not limited to the system vocabulary. STD systems must cope with these so-called *out-of-vocabulary (OOV)* words. Some words are OOV simply because the system vocabulary has a fixed size, whereas others arise from the dynamics of human language. One estimate is that about 20,000 new words are coined each year [Watson 2003]. We are particularly interested in this *absolute* OOV phenomenon which not only relates to engineering/technical limitations but also reflects human language dynamics.

In STD, search terms involving OOV words are known as OOV terms; correspondingly, terms involving only those words within the system vocabulary are in-vocabulary (INV) terms. OOV terms present a significant challenge to STD; in real spoken document retrieval systems, 12% of queries are reported to contain OOV terms [Logan et al. 2000; Olsson and Oard 2009]. Since OOV terms usually convey important information, a good solution to OOV term detection is highly desirable for spoken document indexing and retrieval.

The most popular approach to OOV term detection is based on subword units [Szöke et al. 2008; Mamou et al. 2007; Akbacak et al. 2008]. In this approach, subword representations of search terms are searched for within subword lattices that are generated by a subword-based ASR system. The subword representations are usually obtained from letter-to-sound conversion [Torkkola 1993; Deligne et al. 1995; Luk and Damper 1996; Damper and Eastmond 1997; Black et al. 1998; Daelemans et al. 1999; Bisani and Ney 2003; Taylor 2005]. Among various subword units, phonemes are the most simple and widely used.

Whilst the subword approach enables OOV term detection, performance is always inferior to that for INV terms. One of the principal reasons, we hypothesize, is that special properties of OOV terms are seldom taken into account, and OOV terms are actually treated no differently from INV terms except that the pronunciations are obtained by letter-to-sound mapping rather than from dictionaries. This is clearly suboptimal. A reasonable assumption, is therefore that OOV detection can be improved by considering OOV special properties in term search and confidence estimation.

A wide range of properties can be enumerated to be specific to OOV terms; we consider the following three to be the most relevant for STD:

— Pronunciation uncertainty

Pronunciation uncertainty, although ubiquitous in human speech for any term, is more serious for OOV terms than for INV terms. Firstly, correct pronunciation of OOV terms is unknown, and thus must be predicted with some letter-to-sound approaches. Unfortunately, all the letter-to-sound approaches reported so far suffer from a high word error rate[1], typically in the order of 30-40%, or a phone error rate in the order of 10% [Torkkola 1993; Deligne et al. 1995; Luk and Damper 1996; Damper and Eastmond 1997; Black et al. 1998; Daelemans et al. 1999; Bisani and Ney 2003; Taylor 2005]. The error-prone pronunciation prediction leaves the OOV detection to rely on unreliable lexical forms. Furthermore, the pronunciation of OOV terms exhibits more variation than those of INV terms. When encountered with an unfamiliar word, people tend to slow down, examine the spelling structure, guess the pronunciation, hesitate, and finally try to pronounce. This guess-and-trial process leads to more spontaneous speech phenomena and higher acoustic variation. Importantly, different people might reach different lexical forms by guessing, leading to pronunciation variation at the lexical level. The interweaving of variations at the acoustic and the lexical levels makes OOV treatment highly complicated. Readers are encouraged to refer to [Wang et al. 2010] for further discussion.

— Property diversity
Different OOV terms tend to possess different properties from various aspects, e.g., occurrence rate, phonemic structure, linguistic background, morphological form, etc. This is particularly evident for newly coined terms in the backdrop of increasing international communication and multi-cultural integration in modern society. For instance, some 'natively' created new terms follow English pronunciation/spelling rules strictly, e.g., *GOOGLE*, while some new terms borrowed from other languages do not adhere to the rules, e.g., *KUWAIT*. Also, some new terms tend to remain as jargon limited to a small community, e.g., *ANTIHISTAMINE*, while others obtain quick popularity. Even in those popular terms, some sustain usage over time e.g., *DNA*, while others disappear gradually, e.g., *SARS*. This diversity on the one hand reflects the tendency towards more complexity in human language development; on the other hand, it poses problems in designing a detection scheme widely applicable to all OOV terms in STD systems. Further discussion can be found in [Wang et al. 2011].

— Modeling weakness
OOV terms tend to be weakly modeled by acoustic models (AMs) and language models (LMs) since they have no representation in the training data. Even with subword approaches, the triphone count for AM training and n-gram count for LM training of OOV terms tend to be less than those of INV terms, leading to weak models. We discuss the modeling weakness in more detail in the next section. The consequence of weak models is that lattices generated by the recognition system tend to miss representations (e.g., phoneme sequences) of OOV terms, and the confidence measures derived from AM and LM scores tend to be unreliable.

In previous work, we proposed a stochastic pronunciation modeling (SPM) technique to address pronunciation uncertainty [Wang et al. 2009; 2010], and a discriminative confidence normalization technique to deal with property diversity [Wang et al. 2009]. In this paper, we investigate the modeling weakness and propose a direct posterior confidence estimation to tackle this problem. Instead of being derived from acoustic and language models that are usually less representative of OOV terms, the new confidence measure derives the posterior probability of Eq. 2 'directly' from a discrimina-

[1] In a letter-to-sound task, the word error rate is defined as the proportion of words that are correctly predicted given a test word list.

tive model, e.g. a multi-layer perceptron (MLP). Derived from a wide acoustic context, the new confidence measure ameliorates the weaknesses of acoustic modeling for OOV terms. Furthermore, by localizing at the speech segments of hypothesized detections, this approach avoids long-span linguistic context which is usually problematic for OOV term detection. The direct posterior approach was first presented by the authors as an alternative confidence measure for STD in general [Wang et al. 2009]. In [Tejedor et al. 2009], the direct posterior approach is extended to combine and hybridize heterogeneous STD systems. In [Wang et al. 2010] we propose the idea of applying direct posterior confidence to improve OOV detection and present preliminary results which show that this new confidence measure does provide significant performance enhancement for OOV terms while being less helpful for INV terms. The contribution of the work presented in this paper is two-fold: on the one hand we present an extensive discussion of the modeling weakness associated with OOV terms and thoroughly analyze the ability of direct posterior confidence in addressing the problem; on the other hand, we substantially extend the experiments in [Wang et al. 2010] and provide a comprehensive study for the direct posterior technique as well as its combination with other OOV-oriented approaches such as SPM and discriminative confidence normalization.

The structure of this paper is as follows: in the next section we first review some related work, and then in Section 3 we focus on confidence estimation and present a statistical study to highlight the problem of modeling weakness with OOV terms. In Section 4, we propose the direct posterior confidence approach, including an MLP-based acoustic posterior confidence estimation and an evidence-based confidence integration. The integration of this approach with stochastic pronunciation modeling and discriminative confidence normalization is then presented in Section 5. The experimental setup and evaluation results are reported in Section 6 followed by some discussion in Section 7. Section 8 concludes this work and presents some ideas for future work.

2. RELATED WORK

The study in this paper relates to several research topics including speech recognition, spoken term detection, discriminative modeling and even language development. In this section, we review some related work and focus on OOV treatment and the measurement of phone posterior confidence where this paper makes its contribution.

2.1. Related work on OOV treatment

The most popular approach to detecting OOV terms is based on subword units, particularly phones. The approach was first studied in spoken document retrieval (SDR) [Schäuble and Wechsler 1995; Witbrock and Hauptmann 1997; Wechsler et al. 1998; Ng 2000; Cardillo et al. 2002; Ma and Li 2005; Itoh et al. 2006; Ng 1998], and has been adopted by many researchers in STD, e.g., Szöke et al. [2006], Wallace et al. [2007], Parlak and Saraçlar [2008]. Various other subword units besides phones were also studied in both SDR and STD, e.g., word-fragments [Seide et al. 2004], particles [Logan et al. 2002; Logan et al. 2005], acoustic words [Ma and Li 2005], graphones [Vergyri et al. 2007; Akbacak et al. 2008], multigrams [Pinto et al. 2008; Szöke et al. 2008], syllables [Meng et al. 2007], and graphemes [Wang et al. 2008]. To take advantage of both the word and subword based approaches, combination systems have been proposed in a multitude of research, e.g., [James and Young 1994; James 1996; Jones et al. 1996; Saraclar and Sproat 2004; Szöke et al. 2006; Parlak and Saraçlar 2008; Iwata et al. 2008; Olsson and Oard 2009]. Hybrid approaches which fuse word and subword approaches at the lattice level [Yu and Seide 2004; Meng et al. 2008] or lexicon level [Yazgan and Saraclar 2004; Akbacak et al. 2008; Szöke et al. 2008] have also been proposed.

Some research has been undertaken to tackle OOV pronunciation variation and uncertainty. One popular approach referred to as soft match, allows mismatch in term search subject to a mismatch penalty based on edit distance [James and Young 1994; Thambiratnam and Sridharan 2005; Itoh et al. 2006; Miller et al. 2007] or acoustic confusion [Wechsler et al. 1998; Srinivasan and Petkovic 2000; Szöke et al. 2005; Audhkhasi and Verma 2007; Pinto et al. 2008]. Another well known approach to deal with OOV pronunciation uncertainty is phonetic query expansion which uses n-best pronunciations for OOV term search [Chen 2003; Mamou and Ramabhadran 2008; Sproat et al. 2008; Can et al. 2009]. This is in principle, similar to SPM [Wang et al. 2009] where a joint-multigram model is applied to permit variable-sized grapheme-phoneme correspondence.

Finally, in order to handle property diversity among OOV terms, Miller et al. [2007] propose a term-specific threshold approach which normalizes the diversity in occurrence rate among OOV terms. This approach was adopted by many researchers including Vergyri et al. [2007; Parlak and Saraçlar [2008] and was extended in our study [Wang et al. 2009; Wang et al. 2011] to normalize term-dependent properties with discriminative models.

2.2. Related work on phone posterior confidence estimation

Confidence measures based on phone posterior probabilities have been successfully applied to speech recognition. In [Zavaliagkos et al. 1994], segment phone posterior probabilities generated by various neural networks are used to score ASR hypotheses. Rivlin et al. [1996] use aggregated phone posteriors for utterance rejection, where the frame-level phone posterior probability is calculated from phone class-conditional probabilities using the Bayesian formula. Abdou and Scordilis [2004] propose a similar approach for speech transcription, in which the confidence scores of partial hypotheses in decoding are evaluated and fed back to guide further decoding. An aggregation approach was followed by Bernardis and Bourlard [1998] in the HMM/ANN hybrid framework. Instead of applying the Bayesian formulation, they calculate the frame-level phone posterior probabilities through a neural network and aggregate them to obtain word/utterance-level confidence measures. This approach was adopted by a number of researchers, e.g., Williams and Renals [1999], Silaghi and Bourlard [1999] and Ketabdar et al. [2006]. All the above work concerns speech transcription. For STD, the authors first introduced the MLP-based confidence approach in [Wang et al. 2009] and developed direct posterior-based hybridization and combination in [Tejedor et al. 2009], however the potential of this approach for OOV STD was not investigated until [Wang et al. 2010]. In this paper we extend the discussion in [Wang et al. 2010] and present further evidence to show the effectiveness of direct posterior approach in dealing with modeling weakness associated with OOV terms.

3. PRELIMINARY DISCUSSION

Before presenting the new confidence estimation approach, we give a brief review of various confidence estimation techniques, particularly focusing on the most popular lattice-based confidence estimation. The modeling weakness on OOV terms with this confidence measure is then studied, which motives the idea of the novel direct posterior confidence estimation.

3.1. Confidence estimation for STD

Confidence estimation plays an important role for STD in determining the reliability of putative detections and filtering out false detections. The following review summarizes various approaches to confidence measurement. While we concentrate on STD, we also

look at some related work in speech transcription. Reviews on this subject can also be found in [Siu and Gish 1999; Wessel et al. 2001; Jiang 2005].

3.1.1. Feature-based confidence. The first approach to estimate the confidence is based on some features that are generated during recognition. For example, Rohlicek et al. [1989] proposed the use of duration-normalized acoustic likelihood, Cox and Rose [1996] studied second-phone-recognition normalized acoustic likelihood, Bergen and Ward [1997] used senone-score-normalized acoustic likelihood. Kemp and Schaaf [1997] proposed to use various statistics from lattices, such as link probability, acoustic stability and hypotheses density. Manos and Zue [1997] studied and combined 5 features in a segment-based system, such as segment phonemic match score, lexical weight, etc.

Various models have been used to integrate diverse features into confidence scores. These include decision trees, general linear models, generalized additive models and MLPs [Chase 1997; Gillick et al. 1997; Zhang and Rudnicky 2001]. All this research confirms that the features derived from decoding with suitable normalization and combination, can serve as a good measure for the confidence of a recognition hypothesis or a putative detection of a spoken term.

3.1.2. Likelihood ratio-based confidence. In this approach, the hit/FA decision is cast as testing the null hypothesis that 'the detected term is K' versus the alternative hypothesis that 'the detected term is not K' given the input speech. The decision is made by fixing a threshold on the likelihood ratio of the null hypothesis and the alternative hypothesis [Rahim et al. 1995; Rose et al. 1995; Kamppari and Hazen 2000]. This approach is mainly used in utterance verification [Rahim et al. 1997], in which discriminative training targeted to a minimum classification error rate [Rahim et al. 1995; 1997] or a minimum verification error rate [Sukkar et al. 1996; Sukkar and Lee 1996; Setlur et al. 1996; Sukkar 1998] is widely used.

3.1.3. Discriminative confidence. The third approach to confidence estimation for both speech transcription and keyword spotting/spoken term detection casts the decision making task to a binary classification problem. According to decision theory, an optimal classification strategy should be based on class posterior probabilities, which leads to a discriminative confidence measurement. The first implementation of the discriminative confidence estimation is a *two-class* approach [Young 1994], which models the class-conditional confidence distributions for correct and incorrect detections, and then derives the discriminative confidence using the Bayesian formulation. Jeanrenaud et al. [1995] and Junkawitsch et al. [1996] modeled class-conditional probability density functions and derived term-specific thresholds from them. Fetter et al. [1996] proposed a similar approach, and investigated the relationship between discriminative confidence and likelihood ratio. Note that this two-class approach builds distributions of scores for correct and incorrect recognitions/detections, and that different terms may share the same distributions; this is clearly different from the two-class modeling in the likelihood ratio-based approach where the null-alternative models are built on speech signals and different terms have distinct null-alternative pairs.

The two-class approach requires a model of class-conditional probability distributions and hence is a generative approach; instead, a discriminative approach builds a discriminative model that estimates the classification posterior probabilities directly. Mathan and Miclet [1991] utilized an MLP to generate a discriminative confidence for extraneous speech input rejection. Weintraub et al. [1997] and Vergyri et al. [2006] proposed the same MLP-based confidence to justify recognition hypotheses. Linear discriminative functions were studied by Sukkar and Wilpon [1993], Gillick et al. [1997] and Kamppari and Hazen [2000]. General linear models and generalized additive mod-

els were studied by Siu et al. [1997]. Decision trees were studied by Neti et al. [1997] and Hauptmann et al. [1998], and support vector machines (SVMs) were studied by Zhang and Rudnicky [2001], Sudoh et al. [2006] and Shafran et al. [2006]. Model comparison has been investigated by a number of researchers, e.g., Chase [1997], Schaaf and Kemp [1997], Ábrego [2000] and Zhang and Rudnicky [2001]. Although it is difficult to identify the best model, MLPs and SVMs are generally regarded as the state-of-the-art.

3.1.4. Posterior probability-based confidence. All the above confidence estimation approaches involve a two-step process: in the first step, some informative features are collected, and in the second step these features are combined in a certain way to give a confidence measure. The combination form in the second step is chosen arbitrarily. A more theoretically-sound approach is based on posterior probabilities. Basically, assuming that we know nothing more than the audio stream O, and are given the task to evaluate the confidence that K appears in a particular speech segment, a natural choice for the confidence measure is the posterior probability $P(K|O)$, which we call a-posterior probability-based confidence and has been defined in Eq. 2. According to Bayesian decision theory, a decision based on this posterior probability gives minimum risk when determining which term is contained in the speech. Therefore, the posterior probability is an ideal measure of the confidence of a detection. In practice, the Bayesian rule is usually applied to decompose the posterior probability into a product of the prior probability of the search term and the conditional probability of the speech segment given the term, so we have,

$$\begin{aligned} c &= P(K|O) \\ &= \frac{p(O|K)P(K)}{p(O)}. \end{aligned} \tag{3}$$

Note that Eq. 3 can be regarded as a normalization that amends the likelihood-based confidence to make it comparable across utterances. Rose and Paul [1990] and James [1996] proposed the use of a background model for normalization, and Weintraub [1995] and Jeanrenaud et al. [1995] presented an n-best approach, where the acoustic scores of the hypotheses in the n-best list involving the keyword are accumulated, and then normalized by the summation of the scores of all the n-best hypotheses. Setlur et al. [1996] simplified the n-best approach into a likelihood ratio of the best and the second-best hypotheses, which is a special case of the n-best approach.

3.1.5. Lattice-based confidence. The n-best based confidence was extended to a lattice-based confidence by [Wessel et al. 1998]. A lattice is an acyclic graph in which each node represents a recognition unit (e.g., word or phoneme) and each arc represents a transition from one unit to the next. An arc is associated with relevant information including duration, acoustic and language model scores. A lattice can be regarded as a compact representation of the original speech signal produced by pruning the hypotheses space with the speech recognizer. The *lattice-based confidence* is the posterior probability of the search term appearing in the lattice, and is implemented as the ratio of the score accumulated over all complete paths passing the arcs of the detection to the score accumulated over all complete paths in the lattice. According to Eq. 3, this can be formulated as follows:

$$c_{lat} = \frac{\sum_{\pi_\alpha,\pi_\beta} p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta) P(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)}{\sum_\xi p(O|\xi)P(\xi)} \tag{4}$$

where π_α and π_β denote any path before and after K, with π_α starting from the beginning of the speech and π_β finishing at the end; ξ denotes any complete path through the lattice. Note that $p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)$ corresponds to the acoustic scores that are derived from acoustic models, and $p(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)$ corresponds to the language model scores that are derived from language models. In real implementation, a forward-backward accumulation approach is often used to increase computational efficiency [Szöke et al. 2005]. Because of its theoretical soundness and simplicity in implementation, this lattice-based confidence has been widely used in keyword spotting and STD [Woodland 2000; Szöke et al. 2005; Akbacak et al. 2008; Miller et al. 2007; Mamou et al. 2007; Szöke et al. 2006; Vergyri et al. 2007; Meng et al. 2008]. To distinguish it from other confidence measures that we introduce shortly, the lattice-based confidence is denoted by c_{lat}.

3.2. Modeling weakness of OOV terms

Although the lattice-based confidence measure performs well generally for STD, it is potentially vulnerable to OOV terms. Going back to Eq. 4, we notice that the lattice-based confidence is derived from acoustic and language model scores and is thus heavily dependent on the acoustic and language models of ASR. In the state-of-the-art implementation, they are based on hidden Markov models (HMMs) and n-gram models respectively. Due to the absence of OOV terms in the training corpora, these models tend to be highly biased towards INV terms and are less representative of the OOV terms, leading to the problem of acoustic and language modeling weakness for OOV terms. This weakness on the one hand results in sub-optimal representations of OOV terms in the lattices, and on the other hand leads to unreliable confidence for OOV detections. We investigate the language model weakness and acoustic model weakness in this section.

3.2.1. Language model weakness. The language model weakness for OOV terms is evident. An OOV word that does not exist in the training text is definitely absent in the resulting word-based LM and is hence absent in the lattices generated by the ASR. For phone-based LMs, phone n-grams of the OOV word might exist in the training data, however, they tend to have fewer occurrences compared to the n-grams of INV words, particularly if n is large. In order to verify this conjecture, we count the occurrences of phone n-grams of INV and OOV terms in the training corpus[2] respectively, and compute the average number of occurrences of n-grams with various n. Fig. 2 shows the statistics. As expected, there are fewer occurrences of OOV n-grams than INV n-grams in the training data. Note that the relative difference is higher with larger n-grams, confirming that higher order LMs tend to be weaker for OOV terms.

We can also examine the capability of the resulting phone language models by testing its perplexity on the phone sequences of INV and OOV terms. Fig. 3 shows these results. Here the x-axis represents the LM order n, and the y-axis represents the average perplexity on the phone sequences of INV and OOV terms measured in log scale. It can be seen that with a lower order LM (2-gram, e.g.), the perplexities on INV terms and OOV terms are similar, indicating that the model has identical representative power for both of them; with a higher order language model, however, INV terms and OOV terms exhibit different behaviors. For INV terms, the perplexity consistently decreases with the LM order; for OOV terms, it initially decreases as well, although with a lower rate; when the n-gram order is large, the perplexity however begins to increase. This observation indicates that for INV terms, it is safe to use higher order phone language models to obtain lattices of good quality; for OOV terms, higher order

[2]The details of the term list and the training corpus are presented in Section 6.

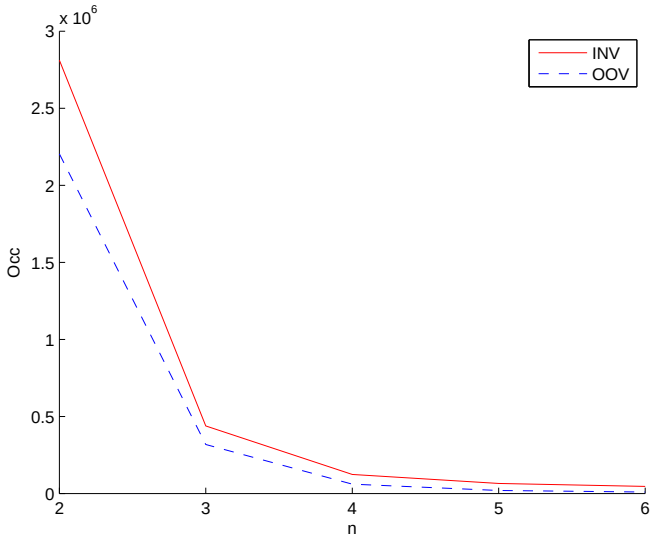

Fig. 2. The average number of training instances of phone n-grams in the training corpus for INV and OOV terms.

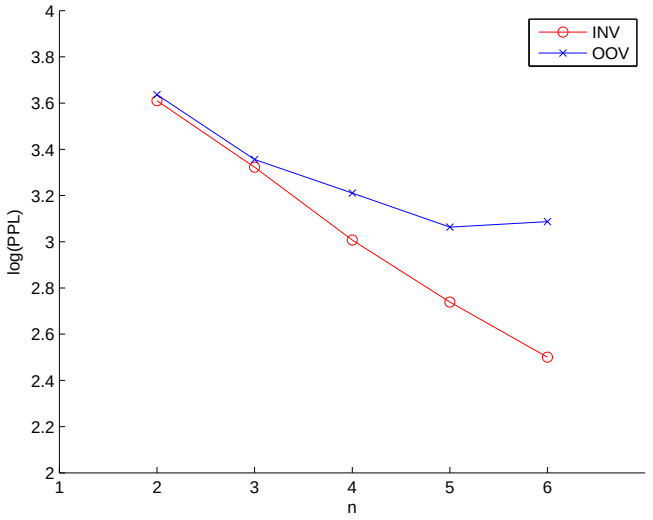

Fig. 3. The log perplexity of LMs of varying order on the phone sequences of INV terms and OOV terms.

language models may result in lattices of lower quality. This is somewhat surprising but still explainable, as the training instances used to generate higher order n-grams are much more representative of INV terms than OOV terms, resulting in models that are strongly biased towards INV terms. Note that the average length of English words is 7.8 (computed from the AMI RT05s dictionary that we used in this study, refer to Section 6), suggesting that an n-gram model whose order is larger than 7 is highly risky for OOV terms, while it is still healthy for INV terms.

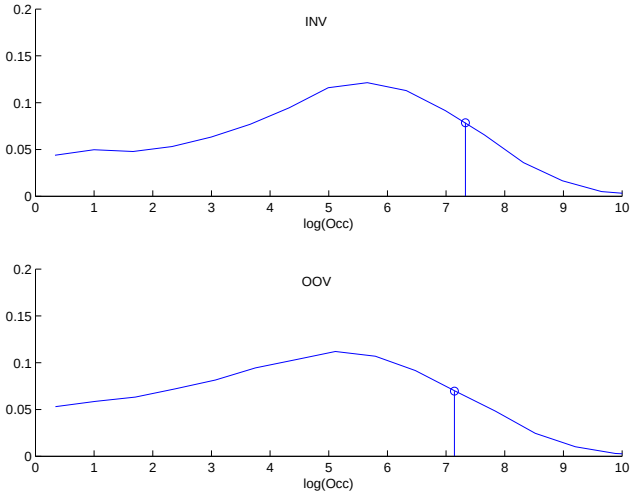

Fig. 4. The distributions of occurrences (in logarithm) of INV triphones and OOV triphones. The vertical lines represent average number of occurrences.

3.2.2. Acoustic model weakness. Similar to the language models, acoustic models also exhibit weakness for OOV terms as a result of fewer training instances. As our ASR component is based on triphone HMMs, we can compare the occurrences of the triphones that are required to represent INV and OOV terms in the training corpus. The statistical distribution of the occurrences of INV and OOV triphones are shown in Fig. 4, with the average number of occurrences being represented by the vertical lines. The result confirms the conjecture that fewer training instances are available for triphones of OOV terms than those of INV terms. However, it also shows that the distributions of INV terms and OOV terms are not significantly different, indicating that the weakness of acoustic models on OOV terms is not as pronounced as that of language models. This is also expected, as the OOV phenomenon relates directly to vocabularies and thus is more linguistic than acoustic in nature. From another perspective, triphones roughly correspond to 3-grams in phone language models, for which, as seen in Fig. 3, there is no substantial difference between INV and OOV terms. If quinphones that corresponds to 5-grams are used, we can expect more significant discrepancy for INV and OOV terms in acoustic modeling.

Finally, it is worth mentioning that the problem of data sparsity for OOV terms can be mitigated by various smoothing techniques, e.g., discounting and back-off in language modeling and tree-based triphone clustering in acoustic modeling; however, these techniques do not eliminate the discrepancy between INV and OOV terms.

3.3. Weakness of lattice-based confidence for OOV terms

The modeling weakness on OOV terms discussed above leads to unreliable acoustic and language model scores, and according to Eq. 4 is inevitably migrated to lattice-based confidence estimation for OOV terms. Moreover, the lattice-based confidence estimation itself is not suitable for OOV terms. From Eq. 4, we see that the lattice-based confidence of a detection is computed from the acoustic and language model scores of the entire utterance. In other words, this is a *global* confidence that takes into account long-span context of the examined detection. Long-span context is usually a desirable

character to consider as the context often conveys valuable information for ascertaining detections. For OOV terms, however, this character might cause serious problems. As we have seen when discussing language model weakness, long-span context may reduce the representative power for OOV terms; this means considering long-span context might introduce unreliable information and hurt the performance for OOV terms.

A possible way to remove the negative impact of long-span context in lattice-based confidence measurement is to diminish the contribution of LM scores in the computation. This is achieved by introducing a LM factor α in the lattice-based confidence estimation, giving

$$c_{lat} = \frac{\sum_{\pi_\alpha,\pi_\beta} p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta) P(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)^\alpha}{\sum_\xi p(O|\xi) P(\xi)^\alpha}.$$

The tuning of the parameter α can be regarded as a trade-off between information and noise introduced due to long-span context. The factor α is chosen to optimize the STD performance on a development set (see Section 6).

This approach can partially mitigate the impact of unreliable long-span context, but does not change the global character of the lattice-based confidence. In fact, even when the LM scores are totally ignored (i.e., $\alpha = 0$), the lattice-based confidence estimation still relies heavily on long-span context due to the use of higher-order language models in lattice generation. Note that choosing a lower-order language model for lattice generation does not work, as INV terms which are the majority in speech require higher-order language models for recognition.

We therefore look for a new confidence estimation, which ideally possesses two properties: first, it should concentrate on speech segment of the detection and ignore the long-span context, i.e., it is a *local* confidence and focuses on the speech segment of the examined detection; second, it is derived from an mechanism different from the conventional acoustic and LM modeling, and thereby more robust for OOV terms. The direct posterior confidence that we introduce in the next section is just such a confidence measure.

4. DIRECT POSTERIOR CONFIDENCE

Although the problem of data sparsity associated with OOV terms always leads to worse performance on OOV terms than on INV terms, an appropriate choice of modeling technique based on various sharing, smoothing and localizing approaches suitable for OOV terms can improve the performance and thus lead to a more robust confidence measurement for OOV terms. In this section we propose such a new confidence estimation approach based on discriminative modeling. Different from the lattice-based confidence that derives the posterior probability $P(K|O)$ from acoustic and language model scores using the Bayesian rule, the new approach derives posterior probabilities from some discriminative models 'directly', and thus can be called the *direct posterior confidence*. In this work we choose an MLP in implementation although any discriminative model may be used.

It is well known that a standard 3-layer MLP network with soft-max output activation can be used to estimate class posterior probabilities. MLPs have been widely used in this fashion for speech recognition. For example, they can be employed to estimate the posterior probabilities for phone classes, given the acoustic features as inputs. These phone posteriors can be used to substitute frame-wise likelihoods of the Gaussian mixture models, resulting in a HMM/ANN hybrid architecture [Morgan and Bourlard 1995]. They can also be combined with conventional acoustic features, e.g., Mel-frequency cepstral coefficients (MFCC), forming the so called 'tandem

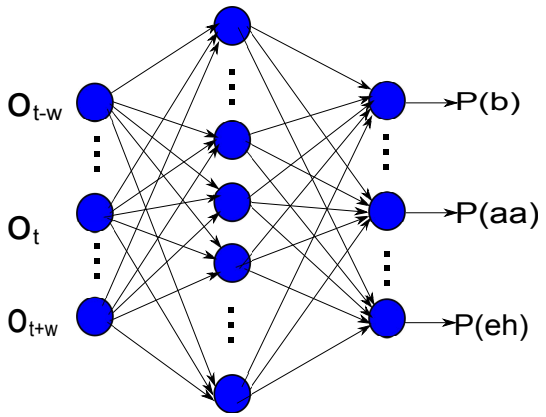

Fig. 5. The MLP structure used to generate frame-wise phone posterior probabilities.

features' that work within the standard HMM-based architecture [Hermansky et al. 2000; Frankel et al. 2008]. All these techniques exhibit considerable advantages. The proposed direct posterior approach follows a similar idea and makes use of phone posteriors to compute local and robust confidence estimates for OOV term detections.

In the rest of this section, we first present the mesh representation of MLP-based phone posteriors, and then propose how to derive acoustic posterior confidence from a mesh. A lattice-based LM posterior confidence is then presented, followed by an evidence-based confidence integration approach which is used to integrate the acoustic posterior confidence and LM posterior confidence.

4.1. Phone posterior mesh

We start with frame-wise phone posterior probability estimation. Let q denote a phone class, the posterior probability of q at time t given the speech O can be written by

$$P(q|O,t) \approx P(q|O_{t-w:t+w}) \tag{5}$$

where $O_{t-w:t+w}$ denotes a windowed speech segment of $2w+1$ frames centered at time t. Here we assume the phone class is dependent only on a local context of the examined frame but independent of its long-span context. This phone posterior, which purely depends on a fixed number of acoustic frames, can be derived from a discriminative model. As in the tandem approach [Hermansky et al. 2000; Frankel et al. 2008], we choose an MLP as the model and a 9-frame window ($w = 4$) to excerpt the local context. This 9-frame configuration roughly corresponds to the length of a typical phone. Fig. 5 illustrates the MLP structure we have used, which involves one hidden layer with sigmoid activation, and soft-max applied on the output. This structure ensures that the output approximates phone posterior probabilities, and the approximation becomes sufficiently robust with adequate training data.

For each frame of the speech, a vector of posterior probabilities can be generated, with each component corresponding to a particular phone class. This forms a phone posterior mesh as illustrated in Fig. 6 where the x-axis represents time and y-axis represents the phone class, and the gray level at point (t,q) in the mesh is proportional to the phone posterior $P(q|O,t)$ given by Eq. 5. We can see some 'phone traces' in the mesh that correspond to the true utterance and its close pronunciations.

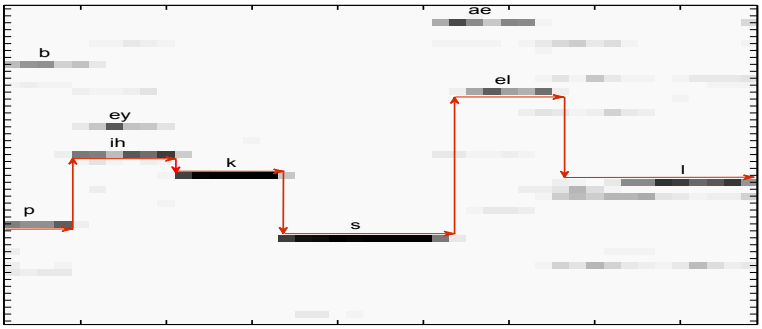

Fig. 6. The posterior mesh generated from MLP.

4.2. Acoustic posterior confidence

With the phone posterior mesh, the posterior probability of a detection d can be computed simply by accumulating the phone posteriors according to its phone trace in the mesh. As an example, Fig. 6 shows how a detection of the term ***pixel*** is handled: from the term detector we know the phone sequence Q and its time stamp, according to which the confidence can be estimated by multiplying all the phone posteriors following the trace ***p ih k s ei l*** illustrated by the solid line. This can be formally written as

$$c_{mlp}(d) = \prod_{t=t_s}^{t_e} P(Q_t|O_{t-w:t+w})$$

where t_s and t_e are the starting and ending time of the detection d respectively, and the notation c_{mlp} is introduced to emphasize that it is derived from MLP-based phone posteriors. Note that the MLP can be substituted by any other discriminative model. In practice, confidence scores of detections of various lengths need to be normalized when compared with each other. A simple way is to average the accumulated confidence over the number of frames, i.e.,

$$c_{mlp}(d) = \left\{ \prod_{t=t_s}^{t_e} P(Q_t|O_{t-w:t+w}) \right\}^{1/(t_e-t_s)}. \quad (6)$$

Compared to the lattice-based confidence (Eq. 4), the MLP-based confidence derives posterior probabilities from a discriminative model (MLP here) instead of generative models (HMMs and n-grams) resorting to the Bayesian rule, which results in confidence scores that are based on entirely different modeling approach and possess distinct properties.

First, MLPs take a different approach in acoustic and phone context modeling. In the lattice-based confidence, the acoustic context (dependence among speech frames) is largely ignored and the phone context is modeled by context-dependent HMMs, i.e., triphone models; In contrast, the MLP model ignores the phonetic context but models a wide range of acoustic context. In the case of OOV terms where phonetic context is less reliable, the acoustically-rich MLP models tend to result in more accurate acoustic modeling than HMMs and therefore provide more reliable confidence estimates.

Second, the MLP-based confidence is local. Although a wide range of acoustic context (9 frames here) is considered, the MLP-based posterior probability is still highly concentrated on the examined frame, leading to a confidence measure concentrated on

the hypothesized detection. As we have discussed in Section 3.3, this local property is particularly desirable for OOV terms where a long-span context may introduce noisy and detracted information for confidence estimation.

Third, the MLP-based posterior confidence is derived purely from acoustic features at the phone level, which is different from the lattice-based confidence where language model scores participate in confidence computation. This means that the impact of LM information can be separated from confidence measurement. For this reason, the MLP-based confidence can also be called the *acoustic posterior confidence*. We shortly present a pure LM-based confidence and show how the acoustic and LM-based confidence can be combined. Even with combination, the acoustic confidence and LM confidence are still well separated as the two confidence estimates are derived from entirely different models. This acoustic-LM separation in confidence estimation is highly desirable for OOV term detection, since the LM information might be less useful or even harmful for OOV terms, as we have seen in Section 3.3.

Last but not the least, MLPs are highly compact as compared to HMMs [Bishop 2006] and the model parameters are globally shared among all phones. This leads to a particular advantage for OOV terms which suffer from limited training data.

4.3. Bayesian acoustic posterior confidence

An obvious limitation of measuring direct posterior confidence is that a phone alignment is required for the computation according to Eq. 6. With a phone-based STD system, the phone alignment can be simply obtained from the term detector; for a general term detector, however, it is in general unavailable. For example, with a word-based STD system, phone alignments cannot be obtained from word lattices. To solve this problem, we can treat the phone alignment as a hidden variable, and marginalize it out by considering all possible alignments. This leads to a Bayesian treatment to the acoustic posterior confidence, given by:

$$c_{mlp}(d) = \left\{ \sum_{\xi} \prod_{t=t_s}^{t_e} P(Q_t^{\xi}|\xi, O_{t-w:t+w}) P(\xi|O_{t-w:t+w}) \right\}^{1/(t_e-t_s)} \tag{7}$$

$$\propto \left\{ \sum_{\xi} \prod_{t=t_s}^{t_e} P(Q_t^{\xi}|\xi, O_{t-w:t+w}) \right\}^{1/(t_e-t_s)} \tag{8}$$

where ξ represents any possible phone alignment, and Q_t^{ξ} denotes the phone category at frame t according to phone alignment ξ. A uniform prior probability $P(\xi|O_{t-w:t+w})$ has been assumed when deriving Eq. 8 from Eq. 7.

This Bayesian treatment enables a flexible framework where the direct posterior approach can be applied to measure detections hypothesized by any STD system without knowing its implementation. As an example, Fig. 7 shows a hybrid architecture where the term detector is based on graphemes and the confidence estimation is based on phone posteriors. The confidence estimation does not care how a detection is hypothesized; instead, it just requires the time stamps t_s and t_e of the assumed detection, and computes the acoustic confidence according to Eq. 8 with a provided phone-based dictionary. Similarly, we can also build a grapheme-based confidence estimation and hybridize it with a phone-based term detector. In general, any type of term detector (word-based/subword-based, n-best-based/lattice-based, etc.) can be hybridized with a direct posterior confidence estimation that can be based on any subword units. In addition, the direct posterior approach also provides a simple combination method where

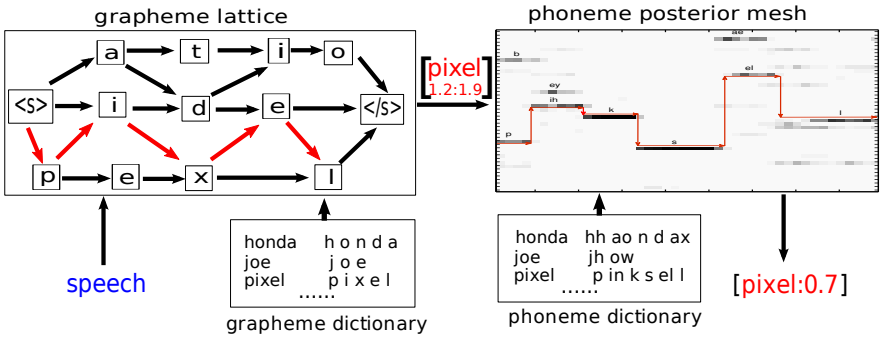

Fig. 7. A hybrid STD architecture involves a grapheme-based term detector and a phone-based confidence estimation.

detections hypothesized by individual STD systems are combined and overlapped detections are merged by accumulating their acoustic posterior confidences. For details on direct posterior-based hybridization and combination, readers are recommended to refer to [Tejedor et al. 2009].

4.4. LM posterior confidence

A particular characteristic of the MLP-based direct posterior confidence estimation is that only the acoustic cues participate in confidence estimation, and the phone posteriors of any consecutive frames are conditionally independent. This leads to a local confidence measure which, as we have discussed, may effectively remove the negative impact of long-span phone context in the case of OOV term detection; however, it also means that some information from linguistic constraints is ignored. This information is potentially beneficial particularly for the INV terms.

In order to retrieve and safely use the information available in the language models, we consider the evidence that a 'linguistic lattice' provides to a putative detection. Similar to the lattice-based confidence estimation, we examine the posterior probability of the phone string of the search term given the lattice, but without considering the acoustic scores. This posterior probability represents the confidence we have for a detection when we observe the search term appearing within the phonemic context. Eq. 9 – 11 formulate this idea, where L denotes the entire phone lattice, K^l denotes the phonetic form of search term K, and C_{K^l} is the context of K^l.

$$c_{lm}(d) = P(K^l|L) \tag{9}$$

$$= \frac{P(K^l, L)}{P(L)} \tag{10}$$

$$= \frac{\sum_{C_{K^l}} P(K^l, C_{K^l})}{P(L)} \tag{11}$$

where c_{lm} denotes the *LM posterior confidence*, given that $P(K^l|L)$ concerns linguistic constraints only. Alike the lattice-based confidence, the LM posterior confidence is global.

4.5. Confidence integration

The MLP-based acoustic and lattice-based LM posterior confidences relate to different aspects of a detection and can thus be combined to improve accuracy. Assuming that the acoustic-based and language-based confidences are given by two independent tests,

and if we further assume that at least one test signifies a positive detection, then the AM and LM confidences may be combined, or fused, as follows:

$$c_{mlp+lm} = 1 - (1 - c_{mlp})^{\alpha}(1 - c_{lm}) \tag{12}$$

where α is a scale factor, and c_{mlp+lm} is the confidence, which integrates the acoustic posterior confidence (c_{mlp}) and LM posterior confidence (c_{lm}), given by Eq. 6 and Eq. 11 respectively. Note that the LM posterior confidence does not provide any more information than what the LM provides in the lattice-based confidence estimation; it is just a convenient form to fuse with the acoustic posterior confidence.

The same approach can be used to combine the acoustic posterior confidence (c_{mlp}) and the lattice-based confidence (c_{lat}), given by Eq. 6 and Eq. 4 respectively. This gives rise to:

$$c_{mlp+lat} = 1 - (1 - c_{mlp})^{\alpha}(1 - c_{lat}) \tag{13}$$

where $c_{mlp+lat}$ is again the combined confidence. Note that additionally combining the LM posterior does not provide any further advantage as the LM information has been utilized in the lattice-based confidence estimation.

We refer to the fusion approach in Eq. 12 and Eq. 13 as the *evidence-based integration*. Compared to other fusion approaches such as linear interpolation and confidence accumulation, a distinct property of this approach is that the resulting confidence is still normalized, i.e., the fused confidence still reflects posterior probabilities. It has been shown that this fusion approach exhibits advantage for OOV STD, given the independence assumption is satisfied [Wang et al. 2011].

5. INTEGRATED SOLUTION FOR OOV TERM DETECTION

We have so far proposed the direct posterior confidence measurement to solve the problem of modeling weakness with OOV terms in STD. As we have discussed, at least two other challenges remain for OOV term detection: the high degree of pronunciation variability and the high diversity with respect to term properties. In this section we describe how the novel approaches that address these two challenges can be combined with the direct posterior confidence estimation. This leads to an integrated and comprehensive solution for OOV term detection.

5.1. Stochastic pronunciation modeling

Stochastic pronunciation modeling (SPM) [Sproat et al. 2008; Wang et al. 2009] has been proposed to address the high degree of pronunciation variability which is typical with OOV terms. With this approach, the pronunciation of a search term is treated as a hidden variable, and a Bayesian treatment is applied to integrate detections based on all possible pronunciations predicted by a letter-to-sound model. Letting Q denote a pronunciation of a search term K, then a detection d of that term, based on this particular pronunciation, may be denoted by:

$$d = (K, Q, \tau, v_a, v_l, ...)$$

where all other symbols have the same meaning as in Eq. 1. We further define the probability of a pronunciation Q for the term K as a pronunciation confidence c_{pron}, i.e.,

$$c_{pron}(d) = P(Q_d|K_d)$$

where K_d is the search term and Q_d is the detected pronunciation represented by the detection d. The confidence of the detection d is then determined according to some composite function of c_{lat} and c_{pron}:

$$c_{spm}(d) = g(c_{lat}, c_{pron})$$

where g is any composite function and c_{spm} denotes confidence according to the SPM. In the original proposal [Wang et al. 2009], a linear composition was utilized as g.

5.2. Discriminative confidence normalization

The term-dependent confidence discrimination technique [Wang et al. 2009] copes with the high diversity of OOV terms with respect to linguistic properties. Using certain discriminative models, the discriminative normalization approach integrates various informative factors encapsulated in a detection and converts them to a hit/FA classification posterior, which is then used in decision making. The central point here is that some term-dependent properties can be taken as the model input which then contribute to the decision making. This leads to a term-dependent decision where diverse properties of OOV terms are taken into account and compensated for in confidence measurement. Formally this approach can be represented as follows:

$$c^{disc}(d) = f(c_{lat}, R_0, R_1, ...)$$

where f represents the discriminative model, which in our work is either an MLP or an SVM. R_0 and R_1 are two occurrence-derived informative factors introduced in [Wang et al. 2009] and defined as

$$R_0(K) = \frac{\sum_i c_{lat}(d_i^K)}{T}$$

and

$$R_1(K) = \frac{\sum_i (1 - c_{lat}(d_i^K))}{T}$$

where T is the length of the audio stream, and d_i^K denotes the i-th detection of term K. Note that R_0 and R_1 are term-dependent and reflect effective hits and effective FAs respectively.

The discriminative confidence normalization is an extension of the term-specific threshold (TST) technique [Miller et al. 2007] which constructs term-dependent decision making by considering term occurrences in evaluation data, and is actually a linear normalization. The advantage of the discriminative normalization rests in the fact that any term-dependent factor can be integrated in the nonlinear mapping represented by the discriminative model, and furthermore it prevents the failure of TST in the case of biased raw confidences.

5.3. Integrated solution

The three techniques we have proposed so far: the direct posterior confidence, as reported here, the stochastic pronunciation modeling and the discriminative normalization as originally reported in [Wang et al. 2009] and [Wang et al. 2009] respectively, tackle the OOV challenge from different perspectives and address different peculiar properties of OOV terms. Additional gains in performance might thus be expected by combining these techniques into an integrated solution. The overall system is illustrated in Fig. 8 and can be formulated as follows:

$$c^{disc}(d) = f(c_{lat}, c_{pron}, c_{mlp}, R_0, R_1, ...).$$

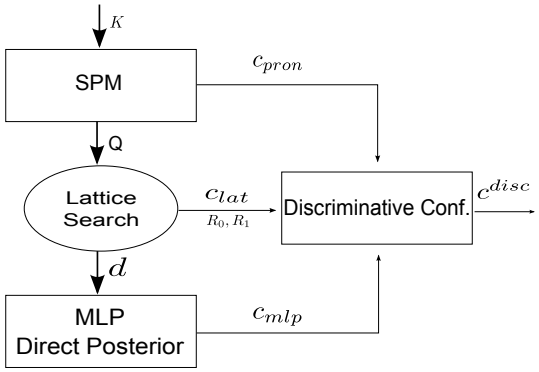

Fig. 8. An illustration of OOV term detection with SPM, discriminative normalization and direct posterior confidence estimation.

Here, according to the SPM, all possible pronunciations are considered in the lattice search, and then each resulting putative detection is assigned a pronunciation confidence c_{pron} given by the letter-to-sound model, a lattice-based detection confidence c_{lat} given by the lattice search, and a direct posterior confidence c_{mlp} given by the MLP-based acoustic posterior prediction. These three confidence estimates, in addition to the term-dependent factors R_0 and R_1 are then fed into the discriminative confidence estimation function f. The resulting discriminative confidence is utilized to make the final hit/FA decision.

6. EVALUATION

6.1. Experimental settings

Our experiments are conducted on English meeting speech recorded from individual headset microphones. Meeting speech is highly spontaneous and presents a substantial challenge to ASR; moreover, speech corpus from meetings tend to involve rich occurrences of OOV terms and thus meet our research objective.

As mentioned before, our research interest focuses on absolute OOV, i.e., genuinely novel terms which are absent not only in the vocabulary but also in the training corpora for AMs and LMs. To create a list of such terms, we compare the AMI dictionary (recently created, in active use and so assumed to represent current usage) and the COMLEX Syntax dictionary v3.1 (published by LDC in 1996 and therefore historical from an STD perspective) and select 412 terms from the AMI dictionary that do not occur in the COMLEX dictionary. These terms to some extent represent the lexical development of English in one decade and thus can be treated as *real* OOV terms. These real OOV terms, however, have only 1143 occurrences in the evaluation data, which is insufficient for a reliable study. In order to overcome this problem, we further add another 70 *artificial* OOV terms to increase the number of OOV occurrences. These artificial terms are all name entities such as person and city names and display more occurrences than most of the real OOV terms. Combing the real and artificial OOV terms results in 482 search terms having a total of 2736 occurrences in the evaluation data. These terms are removed from the system dictionary; furthermore, all utterances and sentences that contain these terms are deleted from the speech and text training corpora. This ensures that they are entirely unseen during system training and parameter tuning and hence comply to the research goal. Besides the OOV terms, 256 INV terms which are mostly person and city names are chosen for the comparative study.

The acoustic models and language models are trained on the same corpora used for model training in the AMI[3] RT05s ASR system [Hain et al. 2006]. After the OOV purge, about 80.2 hours of speech was left for AM training and 521M words of text for language model training. The RT04s development dataset is used for development work. Evaluation corpus comprises the RT04s and RT05s evaluation datasets in addition to a meeting corpus recorded at the University of Edinburgh in 2009 through the AMIDA project. This amounts to 11 hours of speech data for evaluation. For more details of the speech and text corpora, readers are invited to read [Wang 2009].

The acoustic models are 3-state triphone HMMs employing conventional 39 dimensional MFCC features, with cepstral mean and variance normalization (CMN + CVN) applied. The language models are back-off n-gram models with Kneser-Ney discounts and interpolation [Kneser and Ney 1995]. The HTK from Cambridge[4] is used to train the acoustic models and for carrying out decoding; the SRI LM toolkit[5] is used to train n-gram models. Pronunciations of OOV terms are predicted using a letter-to-sound approach based on a joint-multigram model [Deligne et al. 1995; Wang et al. 2009].

Term detection is conducted using the *Lattice2Multigram* tool [Szöke et al. 2005; Szöke et al. 2008] provided by the Speech Processing Group at the Brno University of Technology, with necessary extensions to handle confidence normalization and acoustic posterior estimation. The detection procedure starts with the conversion of search terms into searchable forms which are word and phone sequences for word and phone systems respectively. These forms are organized into a dictionary tree in which any path from the root to a leaf represents a searchable form of the term associated with the leaf. Finally, the occurrences of enquiry terms can be searched for in the lattices, by matching the search forms in the dictionary tree to the partial paths in the lattices. A recursive approach was adopted to conduct this path matching: for each node in the lattice, all the partial paths starting from that node are examined in a depth-first order and only those paths matching a partial path in the dictionary tree are retained and extended. If a leaf node of the dictionary tree is reached, the terms associated with that leaf are detected.

All results reported in this section are those obtained on the evaluation set in terms of average term-weighted value (ATWV) defined by NIST [NIST 2006]. This metric integrates the missing and false alarm probabilities of each term into a single value and then averages over all terms. It is formulated as follows:

$$ATWV = \frac{1}{|\Delta|} \sum_{K \in \Delta} 1 - (P_{miss} + \beta P_{FA}) \tag{14}$$

where Δ denotes the set of search terms and $|\Delta|$ is the number of terms in this set. P_{miss} and P_{FA} are miss probability and false alarm probability respectively and are formally defined as follows:

$$P_{miss} = 1 - \frac{N_{hit}^K}{N_{true}^K}$$

$$P_{FA} = \frac{N_{FA}^K}{T - N_{true}^K}$$

[3]http://www.amiproject.org
[4]http://htk.eng.cam.ac.uk/
[5]http://www-speech.sri.com/projects/srilm/

Table I. Baseline Systems

	ASR			STD (ATWV)	
System	LM	WER/PER	Lattice density	INV	OOV
word	3-gram	39.50%	622	0.5678	-
phone	6-gram	40.49%	805	0.4743	0.2761

where N_{hit}^K and N_{FA}^K represent the number of hits and false alarms of term K respectively, and N_{true}^K is the number of actual occurrences of K in the audio. T denotes the audio length in seconds. $\beta = 999.9$ in (14) is a weight factor to balance the contribution of miss and FA probabilities in the metric. Note that the miss probability is directly related to recall in IR. An obvious property of ATWV is that the metric is term-weighted, which means an STD system cannot obtain good performance by just focusing on frequent terms; instead it has to perform well on all search terms.

In addition to ATWV, NIST proposed the use of detection error trade-off (DET) curves to evaluate system behavior at different hit/FA ratios. Each point in the DET curve represents the recall that the system can achieve with a particular FA probability. The curve is obtained by varying the confidence threshold used for making decisions, and the ATWV result is evaluated at a particular point on the curve. Compared to ATWV which is a point evaluation, DET curves present a global picture of system behavior and thus provide a more systematic understanding. Both ATWV and DET curves are reported in the following experiments.

6.2. Baseline systems

We build two baseline systems, one based on words and the other based on phones. For the word-based system, a 3-gram word language model is used for speech transcription, and for the phone-based system, a 6-gram phone language model is used.[6] For both systems, the term-specific threshold (TST) approach has been applied to conduct confidence normalization. Table I summarizes the characteristics and performance of these two systems, where ASR is evaluated in terms of word error rate (WER) for the word-based system and phone error rate (PER) for the phone-based system, and STD is evaluated in terms of ATWV. The lattice density is computed as the average number of nodes per second, as per the definition in the SRILM toolkit.

We first observe that the word-based system outperforms the phone-based counterpart on INV terms. This is expected as the word-based system uses lexical information which is unavailable for phone-based systems. For OOV terms, the word-based system does not work as no OOV terms appear in the word lattices. The phone-based system can detect part of occurrences of OOV terms, however the performance is significantly deteriorated compared with that on INV terms. With the baseline systems ready, we develop and examine the novel techniques presented in this paper.

6.3. Direct posterior confidence

Applying the direct posterior confidence, the ATWV results are presented in Table II and Table III for the word-based and phone-based systems respectively, where the notation for confidence measures follow the definitions in Section 4. We observe that the acoustic posterior confidence performs worse than the lattice-based confidence for INV terms with the word-based system; when integrated with the LM posterior confidence, however, a significant improvement is obtained ($p < 10^{-5}$ with a t-test). This is consistent with the results on INV terms obtained with the phone-based system, where

[6]We examined various orders of language models, and found that a 6-gram model provides the best ASR performance on the development set. This higher-order language model provides better performance on both INV and OOV STD than other lower order language models.

Table II. Direct Posterior Confidence for Word System

	ATWV	
Confidence	INV	OOV
c_{lat}	0.5678	-
c_{mlp}	0.5605	-
c_{mlp+lm}	0.5894	-
$c_{mlp+lat}$	**0.6134**	-

Table III. Direct Posterior Confidence for Phone System

	ATWV	
Confidence	INV	OOV
c_{lat}	0.4743	0.2761
c_{mlp}	0.4902	0.2971
c_{mlp+lm}	0.4963	0.2941
$c_{mlp+lat}$	**0.5344**	**0.2973**

some performance gains are obtained with the acoustic posterior confidence over the lattice-based confidence, however the improvement is not significant ($p \approx 0.2$); when integrated with the LM posterior confidence, the improvement becomes statistically significant ($p < 10^{-5}$). For OOV terms, the behavior is totally different: the acoustic posterior confidence provides significant performance improvement over the lattice-based confidence ($p < 0.01$), whereas the LM posterior probability does not provide any additional contribution – in fact it deteriorates the performance. This suggests that the LM constraint is informative for INV terms but useless and even harmful for OOV terms. This is consistent with our conjecture that the context information which is captured in context-dependent models in acoustic modeling and n-gram models in language modeling, is not suited to OOV detection; OOV terms are detected more reliably with local confidence with less context interference as observed with the acoustic posterior confidence measure. Finally, combining the acoustic posterior confidence with the lattice-based confidence leads to consistent improvement for both the INV terms and the OOV terms with both the systems, indicating some complementarity between these two approaches which are based on different modeling approaches.

The DET curves, shown in Fig. 9 and Fig. 10, illustrate the differences in detection performance with varying threshold for INV and OOV terms respectively. In each figure, the x-axis represents FA probability and the y-axis represents miss probability. The closer the DET curve is to the origin, the better is the system performance.

We first observe that the INV curves extend to a much lower miss probability (lower right side of the DET plot) than the OOV curves, indicating that much higher precision is obtained on INV terms than on OOV terms. Secondly, we see that the INV curves are almost linear while the OOV curves are concave. This means that for OOV terms, it is rather difficult to get a high recall by just allowing more false alarms, suggesting that the inaccurate speech transcription imposes a strict limitation on the performance of OOV STD.

Concentrating on the DET curves for INV terms (Fig. 9), we see that the acoustic posterior confidence does not give better performance than lattice-based confidence, either with or without the LM posterior confidence. This shows that the lattice-based confidence is good enough for INV term detection and that the new confidence measure does not give much benefit. For OOV terms (Fig. 10), however, we find that the acoustic posterior confidence performs substantially better than the lattice-based confidence, particularly in the region of low false alarms. When integrated with the LM posterior confidence, further gains are obtained, particularly in the low FA area. This is somewhat inconsistent with the ATWV results in Table III, where the LM posterior

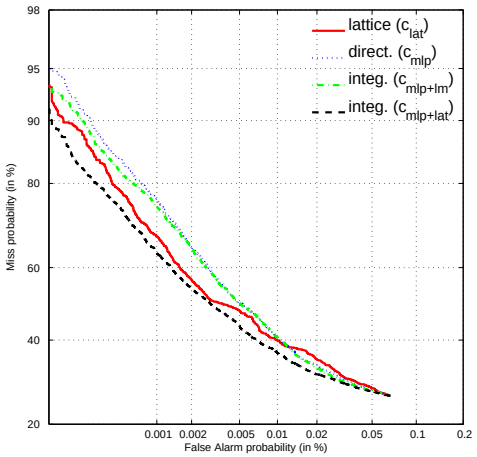

Fig. 9. DET curves on INV terms using various confidence measures.

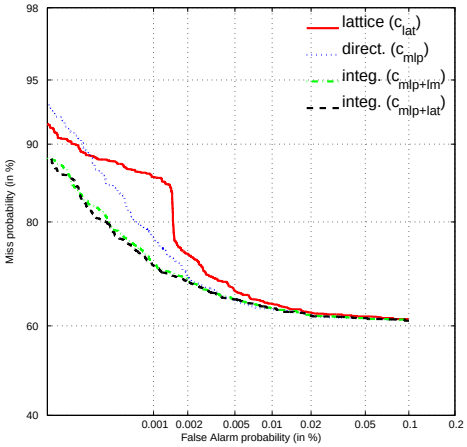

Fig. 10. DET curves on OOV terms using various confidence measures.

confidence contributes very little. This might be due to the fact that the FA suppression is predominantly important in this operating region, so that the linguistic constraint, although noisy, is still beneficial. Nevertheless, the conclusions drawn from the DET curves and the ATWV results are largely consistent: the direct posterior confidence is much more effective than the lattice-based confidence for OOV term detection, and the combination of the two confidences further improves the performance.

6.4. Stochastic pronunciation modeling

As stochastic pronunciation modeling (SPM) applies only to OOV terms, it does not affect the word-based system. Therefore we just present the results on OOV terms with the phone-based system. The experiments are conducted in the same way as in the baseline except that the 1-best pronunciation model is replaced by SPM. The results

Table IV. SPM Results

	ATWV	
Pronunciation Model	INV	OOV
1-best(baseline)	0.4743	0.2761
SPM	-	0.3153

are shown in Table IV, where the first column represents pronunciation modeling approaches that are used: the 1-best system simply takes the best pronunciation, while the SPM-based system considers multiple pronunciations that are generated by a joint multigram model. It can be seen from the results that a substantial performance improvement is achieved with SPM. A pair-wise t-test shows that the improvement is statistically significant ($p < 0.01$).

6.5. Discriminative confidence normalization

We now investigate the contribution of different confidence normalization approaches. The experiments are conducted under the same conditions as in the baseline system except that term-specific threshold (TST) is substituted by discriminative confidence normalization. We examine two discriminative models, an MLP and an SVM. The models are trained as follows. First, STD is carried out on the development set. The resulting detections are accordingly labeled as hits and false alarms and are then employed to train the MLP and the SVM. A 3-layer MLP, whose structure comprises an input layer, a hidden layer with a sigmoid activation and an output layer with a soft-max activation, is trained using the standard error back-propagation algorithm [Bishop 1995]. The number of hidden units, which is chosen to minimize the number of classification errors on the development set by cross-validation, is 30 in our experiments. The SVM is trained with the LIBSVM toolkit [Chang and Lin 2001] with a radial basis kernel function. The parameters, including the error penalty C for classification and the radius scale γ for the kernel, are again optimized by cross-validation, giving $C = 32$ and $\gamma = 0.5$ in our experiments.

Tables V and VI present the results with various confidence normalization techniques, for the word-based and phone-based systems respectively. Compared with TST that is used by the baseline systems, discriminative normalization provides consistent and substantial performance improvements for both the word-based and phone-based systems. Particularly, with the term-dependent quantities R_0 and R_1 involved (see Section 5.2), additional performance gains are obtained. The t-tests show that for INV terms, the improvements are significant ($p < 0.01$) with both word-based and phone-based systems. For OOV terms, discriminative normalization with lattice-based confidence as the only input provides marginally significant improvement over TST ($p \approx 0.05$); with R_0 and R_1 involved, this improvement becomes significant ($p < 0.01$). This conclusion holds no matter which discriminative model is applied, although the SVM exhibits a small advantage with the word-based system while the MLP shows marginal superiority with the phone-based system.

Table V. Confidence Normalization: Word System

		ATWV	
Confidence Normalization	Informative factor	INV	OOV
TST(baseline)	c_{lat}	0.5678	-
MLP	c_{lat}	0.6111	-
MLP	c_{lat}, R_0, R_1	0.6269	-
SVM	c_{lat}	0.6314	-
SVM	c_{lat}, R_0, R_1	**0.6366**	-

Table VI. Confidence Normalization: Phone System

		ATWV	
Confidence Normalization	Informative factor	INV	OOV
TST(baseline)	c_{lat}	0.4743	0.2761
MLP	c_{lat}	0.5453	0.2927
MLP	c_{lat}, R_0, R_1	**0.5460**	**0.2931**
SVM	c_{lat}	0.5432	0.2894
SVM	c_{lat}, R_0, R_1	0.5421	0.2921

Table VII. Direct Posterior with Discriminative Normalization: Word System

		ATWV	
Confidence Normalization	Informative factor	INV	OOV
TST	c_{lat}	0.5678	-
TST	$c_{mlp+lat}$	0.6134	-
MLP	c_{lat}, R_0, R_1	0.6269	-
MLP	$c_{mlp}, c_{lat}, R_0, R_1$	0.6224	-
SVM	c_{lat}, R_0, R_1	**0.6366**	-
SVM	$c_{mlp}, c_{lat}, R_0, R_1$	0.6161	-

Table VIII. Direct Posterior with Discriminative Normalization: Phone System

		ATWV	
Confidence Normalization	Informative factor	INV	OOV
TST	c_{lat}	0.4743	0.2761
TST	$c_{mlp+lat}$	0.5344	0.2973
MLP	c_{lat}, R_0, R_1	**0.5460**	0.2931
MLP	$c_{mlp}, c_{lat}, R_0, R_1$	0.5391	0.3007
SVM	c_{lat}, R_0, R_1	0.5421	0.2921
SVM	$c_{mlp}, c_{lat}, R_0, R_1$	0.5309	**0.3034**

6.6. Direct posterior confidence with discriminative normalization

The direct posterior confidence estimation can be combined with discriminative confidence normalization by extending the input of the discriminative models with acoustic posterior confidence. Table VII and Table VIII present the results of the word-based and phone-based systems respectively. We see that with discriminative confidence normalization applied, involving the acoustic posterior confidence does not provide any benefit for INV terms, with either the word-based or the phone-based system. For OOV terms, however, considerable performance improvement is obtained with the phone-based system. This again supports our conjecture that the direct posterior confidence estimation which aims to ameliorate model weakness along with the lattice-based confidence, is more effective for OOV terms than for INV terms.

6.7. Direct posterior confidence with SPM

Direct posterior confidence estimation can be also integrated with stochastic pronunciation modeling by replacing the lattice-based confidence with the acoustic posterior confidence. The results in such a setting are shown in Table IX. Note that SPM applies to OOV terms only and so just valid for the phone-based system. It can be seen that both SPM and direct posterior confidence estimation improve system performance significantly, and their combination provides additional gains.

6.8. Integrated solution for OOV term detection

Finally we combine the direct posterior confidence measurement, the SPM and the discriminative normalization as an integrated solution. Note that this approach applies to OOV terms and the phoneme-based system only. The ATWV results are shown in Table X. It can be seen that all the three techniques contribute to the OOV term detec-

Table IX. Direct posterior with SPM: Phone System

Pronunciation Model	Confidence	ATWV
1-best	c_{lat}	0.2761
1-best	c_{mlp}	0.2971
SPM	c_{lat}	0.3153
SPM	c_{mlp}	**0.3288**

Table X. Integrated Solution for OOV Terms Detection

		ATWV	
Confidence Normalization	Informative factor	1-best	SPM
TST	c_lat	0.2761	0.3153
TST	c_mlp	0.2971	0.3288
MLP	c_lat, R_0, R_1	0.2931	0.3046
MLP	c_mlp, c_lat, R_0, R_1	0.3007	**0.3423**
SVM	c_lat, R_0, R_1	0.2921	0.3235
SVM	c_mlp, c_lat, R_0, R_1	0.3034	0.3318

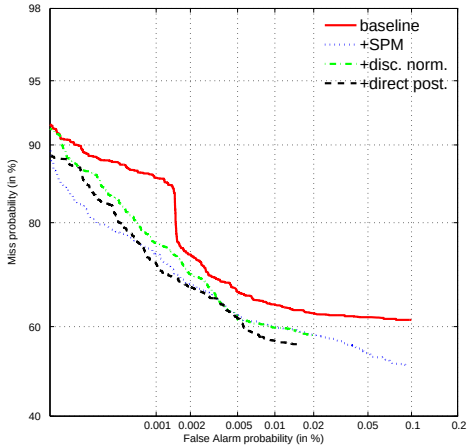

Fig. 11. DET curves for STD systems on OOV terms with the integrated solution.

tion, and their combination (with the MLP-based discriminative normalization) gives the best performance.

DET curves for the integrated system are shown in Fig. 11 where 'disc. norm.' denotes discriminative confidence normalization and 'direct. post.' denotes posterior confidence estimation. For simplicity, only the MLP-based discriminative normalization is shown. They show that the stochastic pronunciation modeling provides the greatest contribution to performance improvement: the DET curve not only falls in the region of lower FA, but also extends to the region of lower miss probability. This means that SPM not only improves detection accuracy, but also improves system recall by considering pronunciation variations. Discriminative normalization does not give much improvement, however it provides a way to integrate various informative factors including the direct posterior confidence. The integration of the three techniques results in the best performance across most of the operating region, but gives relatively poor performance than the SPM-only approach when the FA rate is low. A possible reason is that the MLP model is trained with limited OOV instances, which may lead to unreliable estimation in the region of high precision.

6.9. System combination

In real applications, the primary goal is to obtain the best performance by assembling all available techniques. It is well known that word-based systems outperform phone-based systems when detecting INV terms, while phone-based systems perform better at detecting the OOV terms. Therefore, a commonly used approach to boost STD performance in entirety is to combine these two types of systems. Various combination approaches have been studied, e.g., [James 1996; Jones et al. 1996; Yu and Seide 2004; Szöke et al. 2006; Meng et al. 2008; Akbacak et al. 2008; Olsson and Oard 2009]. In this work we consider a simple linear combination, where the phone-based system is responsible for detecting OOV terms only, and the word-based system works on INV terms.

Given the performance of an STD system on INV and OOV terms, the overall performance of this system is calculated as (15), where $ATWV_{inv}$ and $ATWV_{oov}$ denote the performance on INV and OOV terms respectively, and κ is the OOV rate.

$$ATWV_{overall} = (1 - \kappa) \times ATWV_{inv} + \kappa \times ATWV_{oov} \tag{15}$$

Fig. 12 shows the overall performance of the best word and phone systems in our study, as well as their combination. It shows that when the OOV rate exceeds 18%, the phone-based system outperforms the word-based system; however the combined system always outperforms any individual system.

Note that the above analysis gives only an approximate idea about the potential of the combination approach. In real applications, recognition accuracy and STD performance on INV terms might be highly affected by neighboring OOV terms [Woodland et al. 2000]. In this case the word-based system is not necessarily the best for INV term search and hence linear combination might not necessarily obtain the expected performance. This is particularly important for spoken document retrieval where a query could be as complex as involving multiple INV and OOV terms. More comprehensive combination/hybridization approaches might be helpful, and this is an on-going research topic.

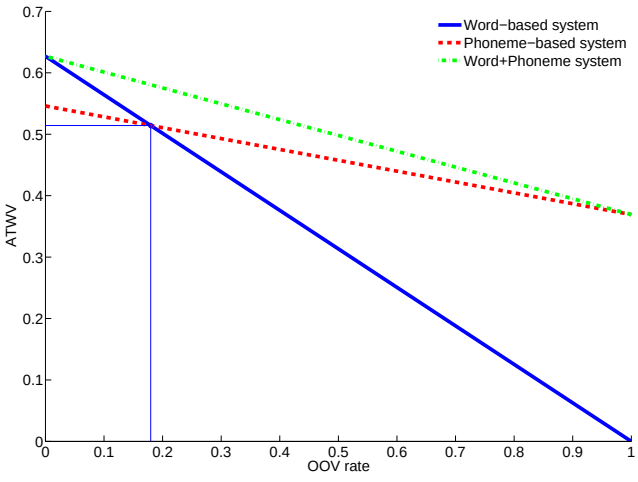

Fig. 12. The overall STD performance of the best word-based and phone-based systems and their combination, with the OOV rate varying from 0.0 to 1.0.

7. DISCUSSION

In this section we compare the proposed direct posterior confidence with two relevant techniques, i.e., discriminative training and confusion networks. These techniques function in a way that is similar to that of the direct posterior confidence to some extent, but are in fact fundamentally different. We argue that the direct posterior approach is more effective for OOV term detection.

First, discriminative model training, e.g., training oriented to minimum mutual information (MMI) [Bahl et al. 1986] or minimum phone errors (MPE) [Povey and Woodland 2002] has been recognized as a standard approach to improve discriminative power of acoustic models in speech recognition. One may argue that the advantage provided by the direct posterior confidence can be assimilated by discriminative training, since a major advantage of the direct posterior approach is that the confidence is derived from discriminative models and thus more discriminative over phone classes. This argument, however, misrepresents the true reasons as to why the direct posterior approach is suited for OOV STD. Certainly the direct posterior confidence is discriminative, but being discriminative on its own is not an advantage over the lattice-based confidence, since the latter is discriminative as well and it works fairly well on INV terms. The main advantage of the direct posterior approach, instead, lies in the fact that rich acoustic context is considered and the problematic long-span linguistic context is avoided. The lattice-based confidence, in contrast, considers limited acoustic context (due to the HMM-based acoustic modeling) and is vulnerable to problematic linguistic context (due to its global property), and is thus less effective for OOV term detection. This weakness associated with the lattice-based confidence can not be addressed by discriminative training although it does help generate lattices of higher quality in general. We therefore argue that the direct posterior approach is more suited for OOV STD than discriminative training, and its contribution can not be assimilated by the latter. In our experiments, the MPE-based training slightly improves ASR performance, however no significant difference is observed for STD performance. This seems consistent with the argument in [Abberley et al. 1998] that for STD, a better detection approach is often more efficient than improving ASR.

Another argument relates to confusion networks (CN) [Mangu et al. 2000]. Within a confusion network, a lattice structure is rearranged to a "sausage" structure by clustering phone arcs with similar time stamps, and the phone posterior of each arc is locally derived from the acoustic and LM scores of all the arcs in its cluster. The confusion network has been used as an alternate representation of lattices for smaller indexes in STD, e.g., [Turunen and Kurimo 2007; Mamou et al. 2007; Parlak and Saraçlar 2008; Can et al. 2009; Natori et al. 2010], and may provide performance similar to that of lattices [Parlak and Saraçlar 2008]. We also recognize that by arc grouping, CN-based confidence may avoid the negative impact of long-span context and may thus be more suited for OOV terms. However, the phone posteriors on the arcs of confusion networks are still derived from the HMM-based acoustic scores, and the disadvantage of weak acoustic context accompanying HMMs still remains. Another disadvantage of confusion networks arises from their approximation to the original lattices, which may in practice reduce performance [Can et al. 2009]. Nevertheless, substituting lattices with confusion networks does simplify the term search and reduces the indexing time.

8. CONCLUSIONS

This paper proposes the use of direct posterior confidence estimates that are derived from an MLP-based phone classifier to tackle the modeling weakness in OOV term detection. Compared to the conventional lattice-based confidence estimation, the new confidence approach considers rich acoustic context but still concentrates on the hy-

pothesized detections. It is therefore better suited to the detection of OOV terms which are usually inadequately represented by acoustic and language models and for which long-span context tends to be problematic. Our experiments, which were conducted on meeting speech which is highly spontaneous and conversational, demonstrate that the direct posterior confidence is more beneficial for OOV terms than for INV terms, and is complementary to the lattice-based confidence. Moreover, results improve significantly when the new confidence measure is integrated with stochastic pronunciation modeling and confidence discrimination, confirming the effectiveness of the integrated solution for OOV term detection. Future work involves investigating other discriminative models such as evolutionary approaches, and exploring confidence enhancement with posterior meshes based on various subword units, as well as more suitable integration approaches.

REFERENCES

ABBERLEY, D., RENALS, S., COOK, G., AND ROBINSON, T. 1998. Retrieval of broadcast news documents with the THISL system. In *Proc. ICASSP'98*. Seattle, Washington, USA, 3781–3784.

ABDOU, S. AND SCORDILIS, M. S. 2004. Beam search pruning in speech recognition using a posterior probability-based confidence measure. *Speech Communication 42,* 3-4, 409–428.

ÁBREGO, G. A. H. 2000. Confidence measures for speech recognition and utterance verification. Ph.D. thesis, University Politècnica de Catalunya.

AKBACAK, M., VERGYRI, D., AND STOLCKE, A. 2008. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. In *Proc. ICASSP'08*. Las Vegas, Nevada, USA, 5240–5243.

AUDHKHASI, K. AND VERMA, A. 2007. Keyword search using modified minimum edit distance measure. In *Proc. ICASSP'07*. Vol. 4. Honolulu, Hawaii, USA, 929–932.

BAHL, L., BROWN, P., DE SOUZA, P., AND MERCER, R. 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proc. ICASSP'86*.

BERGEN, Z. AND WARD, W. 1997. A senone based confidence measure for speech recognition. In *Proc. Eurospeech 97*. Rhodes, Greece, 819–822.

BERNARDIS, G. AND BOURLARD, H. 1998. Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems. In *Proc. ICSLP'98*. Sydney, Australia, 775–778.

BISANI, M. AND NEY, H. 2003. Multigram-based grapheme-to-phoneme conversion for LVCSR. In *Proc. Eurospeech'03*. Geneva, Switzerland, 933–936.

BISHOP, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

BISHOP, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, 225–226.

BLACK, A. W., LENZO, K., AND PAGEL, V. 1998. Issues in building general letter to sound rules. In *Proc. 3rd ESCA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 77–80.

CAN, D., COOPER, E., SETHY, A., WHITE, C., RAMABHADRAN, B., AND SARACLAR, M. 2009. Effect of pronunciations on OOV queries in spoken term detection. In *Proc. ICASSP'09*. Taipei, Taiwan, 3957–3960.

CARDILLO, P. S., CLEMENTS, M., AND MILLER, M. S. 2002. Phonetic searching vs. LVCSR: How to find what you really want in audio archives. *International Journal of Speech Technology 5,* 1, 9–22.

CHAN, C. AND LEE, L. 2010. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. In *Proc. Interspeech'10*.

CHANG, C.-C. AND LIN, C.-J. 2001. *LIBSVM: A library for support vector machines*.

CHASE, L. 1997. Word and acoustic confidence annotation for large vocabulary speech recognition. In *Proc. Eurospeech'97*. Rhodes, Greece, 815–818.

CHEN, S. F. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Proc. Eurospeech'03*. Geneva, Switzerland, 2033–2036.

COX, S. AND ROSE, R. 1996. Confidence measures for the SWITCHBOARD database. In *Proc. ICASSP'96*. Vol. 1. Atlanta, Georgia, USA, 511–514.

DAELEMANS, W., VAN DEN BOSCH, A., AND ZAVREL, J. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning 34,* 1-3, 11–41.

DAMPER, R. AND EASTMOND, J. 1997. Pronunciation by analogy: Impact of implementational choices on performance. *Language and Speech 40,* 1, 1–23.

DELIGNE, S., YVON, F., AND BIMBOT, F. 1995. Variable-length sequence matching for phonetic transcription using joint multigrams. In *Proc. Eurospeech'95*. Madrid, Spain, 2243–2246.

FETTER, P., DANDURAND, F., AND REGEL-BRIETZMANN, P. 1996. Word graph rescoring using confidence measures. In *Proc. ICSLP'96*. Philadelphia, USA, 10–13.

FRANKEL, J., WANG, D., AND KING, S. 2008. Growing bottleneck features for tandem asr. In *Proc. Interspeech'08*. 1549.

GILLICK, L., ITO, Y., AND YOUNG, J. 1997. A probabilistic approach to confidence estimation and evaluation. In *Proc. ICASSP'97*. Munich, Bavaria, Germany, 879–882.

HAIN, T., BURGET, L., DINES, J., GARAU, G., KARAFIAT, M., LINCOLN, M., VEPA, J., AND WAN, V. 2006. The AMI meeting transcription system: Progress and performance. In *Machine Learning for Multimodal Interaction*. Vol. 4299/2006. Springer Berlin/Heidelberg, 419–431.

HAUPTMANN, A. G., JONES, R. E., SEYMORE, K., SLATTERY, S. T., WITBROCK, M. J., AND SIEGLER, M. A. 1998. Experiments in information retrieval from spoken documents. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*. Lansdowne VA, 175–181.

HERMANSKY, H., ELLIS, D. P., AND SHARMA, S. 2000. Tandem connectionist feature extraction for conventional HMM systems. In *Proc. ICASSP'00*. Istanbul, Turkey, 1635–1638.

ITOH, Y., OTAKE, T., IWATA, K., KOJIMA, K., ISHIGAME, M., TANAKA, K., AND WOOK LEE, S. 2006. Two-stage vocabulary-free spoken document retrieval-subword identification and re-recognition of the identified sections. In *Proc. ICSLP'06*. Pittsburgh, USA, 1161–1164.

IWATA, K., SHINODA, K., AND FURUI, S. 2008. Robust spoken term detection using combination of phone-based and word-based recognition. In *Proc. Interspeech'08*. Brisbane, Australia, 2195–2198.

JAMES, D. A. 1996. A system for unrestricted topic retrieval from radio news broadcasts. In *Proc. ICASSP'96*. Vol. 1. Atlanta, Georgia, USA, 279–282.

JAMES, D. A. AND YOUNG, S. J. 1994. A fast lattice-based approach to vocabulary independent wordspotting. In *Proc. ICASSP'94*. Yokohama, Japan, 377–380.

JANSEN, A., CHURCH, K., AND HERMANSKY, H. 2010. Towards spoken term discovery at scale with zero resources. In *Proc. Interspeech'10*.

JEANRENAUD, P., SIU, M., AND GISH, H. 1995. Large vocabulary word scoring as a basis for transcription generation. In *Proc. Eurospeech'95*. Madrid, Spain, 2149–2152.

JIANG, H. 2005. Confidence measures for speech recognition: A survey. *Speech Communication 45,* 4, 455–470.

JONES, G. J. F., FOOTE, J. T., JONES, K. S., AND YOUNG, S. J. 1996. Robust talker-independent audio document retrieval. In *Proc. ICASSP'96*. Atlanta, Georgia, USA, 311–314.

JONES, G. J. F., FOOTE, J. T., SPÄRCK JONES, K., AND YOUNG, S. J. 1996. Retrieving spoken documents by combining multiple index sources. In *Proc. ACM SIGIR'96*. Zurich Switzerland, 30–38.

JUNKAWITSCH, J., NEUBAUER, L., HÖGE, H., AND RUSKE, G. 1996. A new keyword spotting algorithm with pre-calculated optimal thresholds. In *Proc. ICSLP'06*. Pittsburgh, USA, 2067–2070.

KAMPPARI, S. O. AND HAZEN, T. J. 2000. Word and phone level acoustic confidence scoring. In *Proc. ICASSP'00*. Vol. 3. Istanbul, Turkey, 1799–1802.

KANEKO, T. AND AKIBA, T. 2010. Metric subspace indexing for fast spoken term detection. In *Proc. Interspeech'10*.

KEMP, T. AND SCHAAF, T. 1997. Estimating confidence using word lattices. In *Proc. Eurospeech'97*. Rhodes, Greece, 827–830.

KETABDAR, H., VEPA, J., BENGIO, S., AND BOURLARD, H. 2006. Posterior based keyword spotting with a priori thresholds. In *Proc. ICSLP'06*. Pittsburgh, USA, 1642–1645.

KNESER, R. AND NEY, H. 1995. improved backing-off for m-gram language modeling. In *Proc. ICASSP'95*. 181–184.

LEE, H. AND LEE, L. 2010. Integrating recognition and retrieval with user feedback: A new framework for spoken term detection. In *Proc. ICASSP'10*.

LOGAN, B., MORENO, P., AND DESHMUK, O. 2002. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proc. HLT'02*. San Francisco, 31–35.

LOGAN, B., MORENO, P., THONG, J.-M. V., AND WHITTAKER, E. 2000. An experimental study of an audio indexing system for the web. In *Proc. ICSLP'00*. Vol. 2. Beijing, China, 676–679.

LOGAN, B., THONG, J.-M. V., AND MORENO, P. J. 2005. Approaches to reduce the effects of OOV queries on indexed spoken audio. *IEEE Transaction on Multimedia 7,* 5, 899–906.

LUK, R. AND DAMPER, R. 1996. Stochastic phonographic transduction for English. *Computer Speech and Language 10*, 133–153.

MA, B. AND LI, H. 2005. A phonotactic-semantic paradigm for automatic spoken document classification. In *Proc. 28th international ACM SIGIR conference on Research and development in information retrieval*. Salvador, Brazil, 369–376.

MAMOU, J. AND RAMABHADRAN, B. 2008. Phonetic query expansion for spoken document retrieval. In *Proc. Interspeech'08*. Brisbane, Australia, 2106–2109.

MAMOU, J., RAMABHADRAN, B., AND SIOHAN, O. 2007. Vocabulary independent spoken term detection. In *Proc. ACM-SIGIR'07*. Amsterdam, The Netherlands, 615–622.

MANGU, L., BRILL, E., AND STOLCKE, A. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language 14,* 4, 373–400.

MANOS, A. AND ZUE, V. 1997. A segment-based wordspotter using phonetic filler models. In *Proc. ICASSP'97*. Vol. 2. Munich, Bavaria, Germany, 899–902.

MATHAN, L. AND MICLET, L. 1991. Rejection of extraneous input in speech recognition applications using multi-layer perceptrons and the trace of HMMs. In *Proc. ICASSP'91*. Vol. 1. Toronto, Ont., Canada, 93–96.

MENG, S., YU, P., LIU, J., , AND SEIDE, F. 2008. Fusing multiple systems into a compact lattice index for Chinese spoken term detection. In *Proc. ICASSP'08*. Las Vegas, Nevada, USA, 4345–4348.

MENG, S., YU, P., SEIDE, F., AND LIU, J. 2007. A study of lattice-based spoken term detection for Chinese spontaneous speech. In *Proc. ASRU'07*. Kyoto, Japan, 635–640.

MENG, S., ZHANG, W., AND LIU, J. 2010. Combining Chinese spoken term detection systems via side-information conditioned linear logistic regression. In *Proc. Interspeech'10*.

MILLER, D. R. H., KLEBER, M., KAO, C., KIMBALL, O., COLTHURST, T., LOWE, S. A., SCHWARTZ, R. M., AND GISH, H. 2007. Rapid and accurate spoken term detection. In *Proc. Interspeech'07*. Antwerp, Belgium, 314–317.

MORGAN, N. AND BOURLARD, H. 1995. Continuous speech recognition. *IEEE Signal Processing Magazine 12,* 3, 24–42.

MOTLICEK, P., VALENTE, F., AND GARNER, P. 2010. English spoken term detection in multilingual recordings. In *Proc. Interspeech'10*.

NATORI, S., NISHIZAKI, H., AND SEKIGUCHI, Y. 2010. Japanese spoken term detection using syllable transition network derived from multiple speech recognizers' outputs. In *Proc. Interspeech 2010*. Japan.

NETI, C. V., ROUKOS, S., AND EIDE, E. 1997. Word-based confidence measures as a guide for stack search in speech recognition. In *Proc. ICASSP'97*. Munich, Bavaria, Germany, 883–886.

NG, K. 1998. Towards robust methods for spoken document retrieval. In *Proc. ICSLP'98*. Sydney, Australia, 939–942.

NG, K. 2000. Subword-based approaches for spoken document retrieval. Ph.D. thesis, MIT.

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

OLSSON, J. S. AND OARD, D. W. 2009. Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In *SIGIR*. 91–98.

PARADA, C., SETHY, A., DREDZE, M., AND JELINEK, F. 2010. A spoken term detection framework for recovering out-of-vocabulary words using the web. In *Proc. Interspeech'10*.

PARLAK, S. AND SARAÇLAR, M. 2008. Spoken term detection for Turkish broadcast news. In *Proc. ICASSP'08*. Las Vegas, Nevada, USA, 5244–5247.

PINTO, J., SZÖKE, I., PRASANNA, S., AND HEŘMANSKÝ, H. 2008. Fast approximate spoken term detection from sequence of phonemes. In *Proc. The 31st Annual International ACM SIGIR Conference*. Association for Computing Machinery, Singapore, 28–33.

POVEY, D. AND WOODLAND, P. 2002. Minimum phone error and i-smoothing for improved discriminative training. In *Proc. ICASSP'02*. Vol. 1. Orlando, FL, USA, 105–108.

RAHIM, M. G., LEE, C.-H., AND JUANG, B.-H. 1995. Robust utterance verification for connected digits recognition. In *Proc. ICASSP'95*. Vol. 1. Detroit, Michigan, USA, 285–288.

RAHIM, M. G., LEE, C.-H., AND JUANG, B.-H. 1997. Discriminative utterance verification for connected digits recognition. *IEEE Transactions on Speech and Audio Processing 5,* 3, 266–277.

RIVLIN, Z., COHEN, M., ABRASH, V., AND CHUNG, T. 1996. A phone-dependent confidence measure for utterance rejection. In *Proc. ICASSP'96*. Vol. 1. Atlanta, Georgia, USA, 515–517.

ROHLICEK, J. R., RUSSELL, W., ROUKOS, S., AND GISH, H. 1989. Continuous hidden Markov modeling for speaker-independent word spotting. In *Proc. ICASSP'89*. Glasgow, UK, 627–630.

ROSE, R. C., JUANG, B.-H., AND LEE, C.-H. 1995. A training procedure for verifying string hypotheses in continuous speech recognition. In *Proc. ICASSP'95*. Detroit, Michigan, USA, 281–284.

ROSE, R. C. AND PAUL, D. B. 1990. A hidden Markov model based keyword recognition system. In *Proc. ICASSP'90*. Albuquerque, NM, USA, 129–132.

SARACLAR, M. AND SPROAT, R. 2004. Lattice-based search for spoken utterance retrieval. In *Proc. HLT-NAACL 2004*. Boston, USA, 129–136.

SCHAAF, T. AND KEMP, T. 1997. Confidence measures for spontaneous speech recognition. In *Proc. ICASSP'97*. Munich, Bavaria, Germany, 875–878.

SCHÄUBLE, P. AND WECHSLER, M. 1995. First experiences with a system for content based retrieval of information from speech recordings. In *Proc. Workshop on Intelligent Multimedia Information Retrieval (IJCAI'95)*. Montreal, Quebec, Canada, 59–69.

SCHNEIDER, D., MERTENS, T., LARSON, M., AND KOHLER, J. 2010. Contextual verification for open vocabulary spoken term detection. In *Proc. Interspeech'10*.

SEIDE, F., YU, P., MA, C., , AND CHANG, E. 2004. Vocabulary-independent search in spontaneous speech. In *Proc. ICASSP'04*. Vol. 1. Montreal, Quebec, Canada, 253–256.

SETLUR, A. R., SUKKAR, R. A., AND JACOB, J. 1996. Correcting recognition errors via discriminative utterance verification. In *Proc.ICSLP'96*. Philadelphia, USA, 602–605.

SHAFRAN, Z., ROARK, B., AND FISHER, S. 2006. OGI spoken term detection system. In *Proc. NIST spoken term detection workshop (STD 2006)*. Gaithersburg, Maryland, USA.

SILAGHI, M.-C. AND BOURLARD, H. 1999. Iterative posterior-based keyword spotting without filler models. In *Proc. ASRU'99*. Keystone, Colorado.

SIU, M. AND GISH, H. 1999. Evaluation of word confidence for speech recognition systems. *Computer Speech and Language 13,* 4, 299–319.

SIU, M., GISH, H., AND RICHARDSON, F. 1997. Improved estimation, evaluation and applications of confidence measures for speech recognition. In *Proc. Eurospeech'97*. Rhodes, Greece, 831–834.

SPROAT, R., BAKER, J., JANSCHE, M., RAMABHADRAN, B., RILEY, M., SARAÇLAR, M., SETHY, A., WOLFE, P., KHUDANPUR, S., GHOSHAL, A., HOLLINGSHEAD, K., WHITE, C., QIAN, T., COOPER, E., AND ULINSKI, M. 2008. Multilingual spoken term detection: Finding and testing new pronunciations. Tech. rep., JHU.

SRINIVASAN, S. AND PETKOVIC, D. 2000. Phonetic confusion matrix based spoken document retrieval. In *Proc. The 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'00)*. New York, NY, USA, 81–87.

SUDOH, K., TSUKADA, H., AND ISOZAKI, H. 2006. Discriminative named entity recognition of speech data using speech recognition confidence. In *Proc. ICSLP'06*. Pittsburgh, USA, 1153–1156.

SUKKAR, R. A. 1998. Subword-based minimum verification error (SB-MVE) training for task independent utterance verification. In *Proc. ICASSP'98*. Seattle, Washington, USA, 229–232.

SUKKAR, R. A. AND LEE, C.-H. 1996. Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *IEEE Transactions on Speech and Audio Processing 4,* 6, 420–429.

SUKKAR, R. A., SETLUR, A. R., RAHIM, M. G., AND LEE, C.-H. 1996. Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training. In *Proc. ICASSP'96*. Atlanta, Georgia, USA, 518–521.

SUKKAR, R. A. AND WILPON, J. G. 1993. A two pass classifier for utterance rejection in keyword spotting. In *Proc. ICASSP'93*. Vol. 2. Minneapolis, MN, USA, 451–454.

SZÖKE, I., BURGET, L., ČERNOCKÝ, J., AND FAPŠO, M. 2008. Sub-word modeling of out of vocabulary words in spoken term detection. In *Proc. IEEE Workshop on Spoken Language Technology (SLT'08)*. Goa, India, 273–276.

SZÖKE, I., FAPŠO, M., BURGET, L., AND ČERNOCKÝ, J. 2008. Hybrid word-subword decoding for spoken term detection. In *Proc. Speech search workshop at SIGIR (SSCS'08)*. Association for Computing Machinery, Singapore.

SZÖKE, I., FAPŠO, M., KARAFIÁT, M., BURGET, L., GRÉZL, F., SCHWARZ, P., GLEMBEK, O., MATĚJKA, P., KONTÁR, S., AND ČERNOCKÝ, J. 2006. BUT system for NIST STD 2006 - English. In *Proc. NIST Spoken Term Detection Evaluation workshop (STD'06)*. National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

SZÖKE, I., FAPŠO, M., KARAFIÁT, M., BURGET, L., GRÉZL, F., SCHWARZ, P., GLEMBEK, O., MATĚJKA, P., KOPECKÝ, J., AND ČERNOCKÝ, J. 2008. Spoken term detection system based on combination of LVCSR and phonetic search. In *Machine Learning for Multimodal Interaction*. Lecture Notes in Computer Science Series, vol. 4892/2008. Springer Berlin / Heidelberg, 237–247.

SZÖKE, I., SCHWARZ, P., MATĚJKA, P., BURGET, L., KARAFIÁT, M., FAPŠO, M., AND ČERNOCKÝ, J. 2005. Comparison of keyword spotting approaches for informal continuous speech. In *Proc. Interspeech'05*. Lisbon, Portugal, 633–636.

TAYLOR, P. 2005. Hidden Markov models for grapheme to phoneme conversion. In *Proc. Interspeech'05*. Lisbon, Portugal, 1973–1976.

TEJEDOR, J., SZOKE, I., AND FAPSO, M. 2010. Novel methods for query selection and query combination in query-by-example spoken term detection. In *Proc. SSCS'10*.

TEJEDOR, J., WANG, D., KING, S., FRANKEL, J., AND COLÁS, J. 2009. Term-dependent confidence for out-of-vocabulary term detection. In *Proc. Interspeech'09*. Brighton, UK, 2131–2134.

THAMBIRATNAM, K. AND SRIDHARAN, S. 2005. Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting. In *Proc. ICASSP'05*. Vol. 1. Philadelphia, Pennsylvania, USA, 465–468.

TORKKOLA, K. 1993. An efficient way to learn English grapheme-to-phoneme rules automatically. In *Proc. ICASSP'93*. Minneapolis, MN, USA, 199–202.

TURUNEN, V. T. AND KURIMO, M. 2007. Indexing confusion networks for morph-based spoken document retrieval. In *Proc. SIGIR'07*. 631–638.

VERGYRI, D., SHAFRAN, I., STOLCKE, A., GADDE, R. R., AKBACAK, M., ROARK, B., AND WANG, W. 2007. The SRI/OGI 2006 spoken term detection system. In *Proc. Interspeech'07*. Antwerp, Belgium, 2393–2396.

VERGYRI, D., STOLCKE, A., GADDE, R. R., AND WANG, W. 2006. The SRI 2006 spoken term detection system. In *Proc. NIST spoken term detection workshop (STD 2006)*. Gaithersburg, Maryland, USA.

WALLACE, R., VOGT, R., BAKER, B., AND SRIDHARAN, S. 2010. Optimising figure of merit for phonetic spoken term detection. In *Proc. ICASSP'10*.

WALLACE, R., VOGT, R., AND SRIDHARAN, S. 2007. A phonetic search approach to the 2006 NIST spoken term detection evaluation. In *Proc. Interspeech'07*. Antwerp, Belgium, 2385–2388.

WANG, D. 2009. Out-of-vocabulary spoken term detection. Ph.D. thesis, The Center for Speech Technology Research, Edinburgh University.

WANG, D., EVANS, N., KING, S., AND TRONCY, R. 2011. Handling overlaps in spoken term detection. In *Proc. ICASSP'11*. Prague, Czech.

WANG, D., FRANKEL, J., TEJEDOR, J., AND KING, S. 2008. A comparison of phone and grapheme-based spoken term detection. In *Proc. ICASSP'08*. 4969–4972.

WANG, D., KING, S., EVANS, N., FRANKEL, J., AND TRONCY, R. 2010. Direct posterior confidence for out-of-vocabulary spoken term detection. In *Proc. ACM SSCS'10*. Italy.

WANG, D., KING, S., AND FRANKEL, J. 2009. Stochastic pronunciation modelling for spoken term detection. In *Proc. Interspeech'09*. Brighton, UK, 2135–2138.

WANG, D., KING, S., AND FRANKEL, J. 2010. Stochastic pronunciation modeling for out-of-vocabulary spoken term detection. *IEEE Trans. on Audio, Speech, and Language Processing*.

WANG, D., KING, S., FRANKEL, J., AND BELL, P. 2009. Term-dependent confidence for out-of-vocabulary term detection. In *Proc. Interspeech'09*. Brighton, UK, 2139–2142.

WANG, D., TEJEDOR, J., FRANKEL, J., AND KING, S. 2009. Posterior-based confidence measures for spoken term detection. In *Proc. ICASSP'09*. Taiwan, 4889–4892.

WANG, D., TEJEDOR, J., KING, S., AND FRANKEL, J. 2011. *Term-dependent Confidence Normalisation for Out-of-Vocabulary Spoken Term Detection*. EURECOM. submitted to Journal of Computer Science and Technology.

WATSON, D. 2003. *Death Sentence, The Decay of Public Language*. Knopf, Sydney.

WECHSLER, M., MUNTEANU, E., AND SCHÄUBLE, P. 1998. New techniques for open-vocabulary spoken document retrieval. In *Proc. ACM SIGIR 1998*. Melbourne, Australia, 20–27.

WEINTRAUB, M. 1995. LVCSR log-likelihood ratio scoring for keyword spotting. In *Proc. ICASSP'95*. Vol. 1. Detroit, Michigan, USA, 297–300.

WEINTRAUB, M., BEAUFAYS, F., RIVLIN, Z., KONIG, Y., AND STOLCKE, A. 1997. Neural-network based measures of confidence for word recognition. In *Proc. ICASSP'97*. Munich, Bavaria, Germany, 887–890.

WESSEL, F., MACHEREY, K., AND SCHLÜTER, R. 1998. Using word probabilities as confidence measures. In *Proc. ICASSP'98*. Vol. 1. Seattle, Washington, USA, 225–228.

WESSEL, F., SCHLTER, R., MACHEREY, K., AND NEY, H. 2001. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing 9,* 3, 288–298.

WILLIAMS, G. AND RENALS, S. 1999. Confidence measures from local posterior probability estimates. *Computer Speech and Language 13,* 4, 395–411.

WITBROCK, M. J. AND HAUPTMANN, A. G. 1997. Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *Proc. 2nd ACM International conference on Digital Libraries*. Philadelphia PA, USA, 30–35.

WOODLAND, G. E. P. 2000. Large vocabulary decoding and confidence estimation using word posterior probabilities. In *Proc. ICASSP'2000*.

WOODLAND, P., JOHNSON, S. E., JOURLIN, P., AND SPÄRCK JONES, K. 2000. Effects of out of vocabulary words in spoken document retrieval. In *Proc. ACM SIGIR 2000*. Athens, Greece, 372–374.

YAZGAN, A. AND SARACLAR, M. 2004. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. In *Proc. ICASSP'04*. Vol. 1. Montreal, Quebec, Canada, 745–748.

YOUNG, S. R. 1994. Detecting misrecognitions and out-of-vocabulary words. In *Proc. ICASSP'94*. Vol. 2. Adelaide, SA, Australia, 21–24.

YU, P. AND SEIDE, F. 2004. A hybrid word / phoneme-based approach for improved vocabulary-independent search in spontaneous speech. In *Proc. ICSLP'04*. Jeju, Korea, 293–296.

ZAVALIAGKOS, G., ZHAO, Y., SCHWARTZ, R., AND MAKHOUL, J. 1994. A hybrid segmental neural net/hidden Markov model system for continuous speech recognition. *IEEE Transactions on Speech and Aduio Processing 2,* 1, 151–160.

ZHANG, R. AND RUDNICKY, A. I. 2001. Word level confidence annotation using combinations of features. In *Proc. Eurospeech'01*. Aalborg, Denmark, 2105–2108.