Feature analysis for discriminative confidence estimation in spoken term detection

Feature Analysis for Discriminative Confidence Estimation in Spoken Term Detection

Javier Tejedor[a], Doroteo T. Toledano[b], Dong Wang[c], Simon King[d], José Colás[a]

[a]*Human Computer Technology Laboratory, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain*
[b]*ATVS-Biometric Recognition Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain*
[c]*Center for Speech and Language Technologies, Tsinghua University, Beijing 100084, P.R. China*
[d]*The Centre for Speech Technology Research, University of Edinburgh, UK*

Abstract

Discriminative confidence based on multi-layer perceptrons (MLPs) and multiple features has shown significant advantage compared to the widely used lattice-based confidence in spoken term detection (STD). Although the MLP-based framework can handle any features derived from a multitude of sources, choosing all possible features may lead to over complex models and hence less generality. In this paper, we design an extensive set of features and analyze their contribution to STD individually and as a group. The main goal is to choose a small set of features that are sufficiently informative while keeping the model simple and generalizable. We employ two established models to conduct the analysis: one is linear regression which targets for the most relevant features and the other is logistic linear regression which targets for the most discriminative features. We find the most informative features are comprised of those derived from diverse sources (ASR decoding, duration and lexical properties) and the two models deliver highly consistent feature ranks. STD experiments on both English and Spanish data demonstrate significant performance gains with the proposed feature sets.

Email addresses: javier.tejedor@uam.es (Javier Tejedor), doroteo.torre@uam.es (Doroteo T. Toledano), wangdong99@mails.tsinghua.edu.cn (Dong Wang), simon.king@ed.ac.uk (Simon King), jose.colas@uam.es (José Colás)

Keywords: feature analysis, discriminative confidence, spoken term detection, speech recognition.

1. Introduction

1.1. Spoken Term Detection

The enormous amount of speech information now stored in audio repositories motivates the development of automatic audio indexing and spoken document retrieval methods. Spoken Term Detection (STD), defined by NIST as *searching vast, heterogeneous audio archives for occurrences of spoken terms* (NIST, 2006), is a fundamental building block of such systems (Mamou and Ramabhadran, 2008; Can et al., 2009; Vergyri et al., 2007; Akbacak et al., 2008; Szöke et al., 2008b,a; Thambiratmann and Sridharan, 2007; Wallace et al., 2010; Jansen et al., 2010; Parada et al., 2010; Chan and Lee, 2010; Chen et al., 2010; Motlicek et al., 2010), and its development has been strongly influenced by NIST STD evaluations (NIST, 2006, 2013).

The standard STD architecture is comprised of two main stages: indexing by the Automatic Speech Recognition (ASR) subsystem, then search by the STD subsystem, as depicted in Figure 1. The ASR subsystem transforms the input speech into word or sub-word lattices. The STD subsystem comprises a *term detector* and a *decision maker*. The term detector searches for putative occurrences of the query terms in the word/sub-word lattices – it hypothesizes detections – and the decision maker then decides whether each detection is reliable enough to be considered as a hit or should be rejected as a false alarm (FA). A tool provided by NIST is used for performance evaluation. It must be noted that the ASR subsystem must run just once and therefore the STD subsystem cannot make use of the speech signal directly.

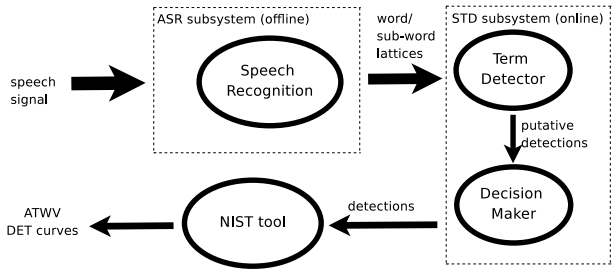

Figure 1: *The standard STD architecture and evaluation.*

Searching the output of a Large Vocabulary Continuous Speech Recognition (LVCSR) system, i.e., word lattices, has been shown to work well when the query terms are only composed of in-vocabulary (INV) words, since these will be in the LVCSR system vocabulary and therefore will occur in the word lattices. However, as noted by Logan et al. (2000), about 12% of users' queries typically contain out-of-vocabulary (OOV) words, which will never be found in the word lattices, because they do not appear in the LVCSR system vocabulary. Common approaches to solve this problem usually involve producing sub-word (typically phone/phoneme) lattices with the ASR subsystem, and then searching for sub-word representations of the enquiry terms (Saraçlar and Sproat, 2004; Mamou et al., 2007; Can et al., 2009; Szöke et al., 2006; Wallace et al., 2007; Parlak and Saraçlar, 2008). Other sub-word units are possible, such as syllables (Meng et al., 2007), graphemes (Wang et al., 2008; Tejedor et al., 2008) or multi-grams (Pinto et al., 2008; Szöke et al., 2008a).

In STD, a *confidence score* is assigned to each putative occurrence detected in the lattice, which reflects the possibility of it being a real occurrence. A widely used confidence score that can be derived from the lattice is defined as follows:

$$c_f = \frac{\sum_{\pi_\alpha,\pi_\beta} p(O|\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta) P(\pi_\alpha, K_{t_s}^{t_e}, \pi_\beta)}{\sum_{\varsigma} p(O|\varsigma) P(\varsigma)} \tag{1}$$

where $K_{t_s}^{t_e}$ denotes a detection of K, which is a partial path that starts at t_s and ends at t_e and corresponds to the pronunciation of term K. c_f is the confidence of $K_{t_s}^{t_e}$. π_α and π_β denote paths before and after K respectively, with π_α starting from the beginning of the audio and π_β ending at the end of the audio. ς in the denominator represents any full path in the lattice. Note that a particular term occurrence may correspond to a group of overlapped detections $\{K_{t_s^i}^{t_e^i}\}$. In that case, the detection group is treated as a single detection, and c_f is derived by a certain merging scheme (Wang et al., 2011). In this work, we simply choose the best confidence of the group members as c_f.

Based on the confidence scores, the decision maker determines which putative occurrences are reliable enough to be called *detections*. If a detection actually appears in the audio, it is called a *hit*. Otherwise, it is called a *false alarm*. Any occurrence of the query term in the audio that is not hypothesized by the STD system is called a *miss*.

To evaluate STD system performance, NIST defines an evaluation metric called *actual term weighted value (ATWV)* (NIST, 2006), which integrates the hit rate and false alarm rate into a single metric and then averages over all search terms:

$$ATWV = \frac{1}{|\Delta|} \sum_{K \in \Delta} \left(\frac{N_{hit}^K}{N_{true}^K} - \beta \frac{N_{FA}^K}{T - N_{true}^K} \right) \tag{2}$$

where Δ denotes the set of search terms and $|\Delta|$ represents the number of terms in this set. 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, and $\beta = 999.9$ is a weight factor.

1.2. Motivation and organization of this paper

Various factors impact the performance of STD systems, such as acoustic properties of speech signals, lexical characteristics of search terms, occurrence rates and positions of terms within the evaluation data, etc. These factors also influence the reliability of detections and hence can be utilized in estimating the confidence of detections. Research has been conducted on confidence estimation utilizing various informative factors, in both automatic speech recognition and keyword spotting, e.g., Rohlicek et al. (1989); Cox and Rose (1996); Bergen and Ward (1997); Kemp and Schaaf (1997); Ou et al. (2001); Ayed et al. (2002); Jiang (2005), and various methods have been employed to combine the heterogeneous informative factors, including decision trees (DT), general linear models (GLMs), generalized additive models (GAMs) and multi-layer perceptrons (MLPs) (Chase, 1997; Gillick et al., 1997; Zhang and Rudnicky, 2001). It has been found that features derived from multiple sources – with appropriate normalization – can be combined to serve as a good measure of confidence, which can in turn be used to evaluate the correctness of a recognition hypothesis or a keyword detection.

In STD, Vergyri et al. (2007) used MLPs to combine various informative factors into a discriminative confidence. We extended this to a discriminative confidence normalization technique (Wang et al., 2009b). This technique provides a general framework which allows any informative factors, or *features*, to be combined and integrated into an unbiased confidence measure for STD. For example, Wang et al. (2009b) used some term-dependent features to compensate for the high diversity among OOV terms, and Tejedor

et al. (2010) extensively studied various prosodic, lexical and duration-based features.

Although the MLP-based discriminative confidence normalization approach provides a general solution for integrating multiple informative factors, and leads to improved performance, the integration itself remains a 'black-box' to a large extent due to the use of MLPs. Contributions from genuinely informative features and trivial features are not separable and so too many features are used because it is not possible to select only the useful ones. Using more features requires more complex model structures, which may reduce the capacity of the model to generalize to new data. This is particularly problematic when the data for training the MLP model are limited; this is the case when the search terms are OOV, since these never occur in training data.

This serves as the motivation for the study on feature analysis presented here. Ideally, we would wish to test the contribution of each potentially-informative feature to overall STD performance, and choose those only that have a significant contribution to discriminative confidence estimation. This not only saves cost in model training and scoring, but more importantly reduces the chance of over-fitting to the training data.

We could use the MLP model itself to analyze and select the features, however, this requires relatively heavy computation; more importantly, using such a complex model as an MLP for feature selection could result in coming to conclusions that do not generalize well. We therefore use simple and well-known models to conduct an analysis on feature relevance; we then use the results of this analysis in an STD system built using MLP-based confidence estimation and normalization.

In previous work (Tejedor et al., 2010), we employed linear regression (LR) to conduct the relevance analysis and feature selection. We found that STD performance could be reduced by using less-relevant features, whilst more informative features generally improved performance. This is consistent with our hypothesis that blindly pooling together both informative and uninformative features may lead to lower model generality and hence performance reduction.

A particular concern with our previous LR analysis, however, is that LR focuses on the most relevant features, while the STD task, which is essentially a classification problem, requires the most *discriminative* features. This raises doubts about the LR-based analysis since the most relevant features are not necessarily the most discriminative ones. Moreover, in our particu-

lar application, LR is employed to predict binary variables (hit/FA). This is essentially a classification task, for which LR is normally considered to be problematic.

We therefore consider the logistic linear regression (LLR) to conduct feature analysis and selection. LLR is an extension of LR, which is more suitable for classification tasks and targets the most discriminative features. An interesting finding is that the relevance-oriented LR analysis and the discrimination-oriented LLR analysis lead to highly consistent feature ranks. This confirms that the simple LR analysis is an effective feature selection approach for STD. Note that some researchers have shown that LR and LLR tend to provide consistent results with abundant training data, which coincides with our findings (Hellevik, 2009).

In addition to this extension of previous work, we also extend our paper in two other directions: we include more features (particularly phone-level features) in the analysis; we study two languages (English and Spanish) in two domains (meetings for English; read speech in Spanish) in order to discover general as well as task-specific features.

We conduct our analysis on OOV terms. The main reason is that STD usually suffers from significant performance degradation on OOV terms, and the unreliable confidence estimation is known to be a major cause (Wang, 2009). We are interested in tackling this difficulty by involving the most discriminative features within the discriminative confidence normalization framework. For INV terms, the standard lattice-based confidence is usually sufficient to obtain a high ATWV, and thus the complicated feature selection is applicable but not essential.

The rest of the paper is organized as follows: in the next section we introduce some background information including a summary of some related work and a short review of the discriminative confidence estimation framework. Then in Section 3 we present the experimental configurations used in this work. Section 4 presents the features considered for study. Individual feature analysis based on histograms, LR and LLR is presented in Section 5. Section 6 presents an incremental feature selection approach based on LR and LLR. Section 7 presents STD experiments based on the two feature selection techniques. The paper is concluded in Section 8 together with some discussion and ideas for future work.

2. Preliminary discussion

The contribution of this paper is to start from a large set of candidate predictive features and find the most informative feature set for STD confidence estimation, based on the discriminative confidence normalization framework. We commence by summarizing some related work, particularly focusing on collecting the candidate feature set, feature selection and feature combination. Given this background knowledge, we then present the discriminative confidence normalization framework on which our analysis is based.

2.1. Related work

The task of STD is related to ASR, information retrieval, statistic modeling, decision theory, etc; a full literature review would therefore be impossible in the scope of this paper. Considering the main research purpose, we focus only on past work related to feature analysis.

2.1.1. Feature collection

Many informative features for confidence estimation have been proposed in the literature. Cox and Rose (1996) studied second-phone-recognition normalized acoustic likelihood, duration, number of phonemes and number of decoding hypotheses in a simple Bayesian classifier as confidence measure in an LVCSR system. Bergen and Ward (1997) used senone-score-normalized acoustic likelihood derived from the word recognizer as a confidence measure for word and phone recognition, and word clustering in a semi-continuous HMM-based recognizer. Kemp and Schaaf (1997) employed various statistics derived from lattices, such as link probability, acoustic stability and hypothesis density as inputs for linear discriminant analysis, neural network and decision tree-based classifiers in an LVCSR system. Chase (1997) proposed features derived from decoding, such as the content of the N-best list, language model score, word pronunciation, word frequency in acoustic training material, separate-phoneme-recognition score, separate-frame-by-frame recognition score, a match count at frame level, a phonologically-based similarity measure and an empirically derived confusion-based distance, senone-based acoustic likelihood normalization, frame-based word duration and number of phones of the word as input features for a post-classifier based on DT, GLMs, GAMs and neural networks for an LVCSR system. Gillick et al. (1997) studied word duration, language model score, acoustic score normalized by the best score, n-best score (as the fraction of the n-best list that

contains the given word in the correct position) and the active node count (as the average number of active states on each frame over a word) as input features for a general linear model in an LVCSR system. The study reported in (Zhang and Rudnicky, 2001) included acoustic features, language model features, word lattice features, N-best features, and parser-based features derived from the language model features and the grammar (parsing-mode and slot-backoff-mode) as input features for three different post-classifiers (DT, neural network and support vector machine (SVM)) in an LVCSR system. Recent work (Goldwater et al., 2009) has proposed disfluency-based features, speaker sex, broad class-based features, turn boundary-based features, language model-based features, pronunciation-based features (word length, number of pronunciations, number of homophones, number of neighbors, and frequency-weighted homophones/neighbors), prosodic features (pitch, intensity, speech rate, duration and log jitter) and concluded that extreme prosodic values, words following a speaker turn and preceding disfluent interruption contribute most to a high word error rate (WER). To the best of our knowledge, there has not been such a systematic analysis of relevant features for STD.

For keyword spotting/spoken term detection, Rohlicek et al. (1989) proposed to use duration-normalized acoustic likelihood as a confidence measure for each keyword in a filler model-based keyword spotting system, and Manos and Zue (1997) studied various features in a filler model-based keyword spotting system to compute the confidence score for each word, which included a segment phonemic match score, a score based on the probability of the particular segmentation, a lexical weight, a phone duration-based score, and a bigram transition score. Ou et al. (2001) employed word posterior likelihoods derived from keyword, anti-keyword and non-keyword models and duration as features in a neural network classifier for utterance verification within a filler model-based keyword spotting system whereas Ayed et al. (2002) used information based on the number of frames, the phone posterior probability, the frame-based phone posterior probability and the duration-based phone posterior probability along with the number of phones as lexical feature in an SVM classifier for utterance verification in an LVCSR-based keyword spotting system.

2.1.2. Feature selection

A straightforward approach to feature selection is based on ranking, where features are selected according to their relevance to some decision target. The

relevance can be computed according to various criteria, such as Pearson correlation coefficient R^2, Fisher's criterion, T-test criterion, classification error rate (CER) with single variable classifiers (Furey et al., 2000; Tusher et al., 2001; Hastie et al., 2001; Forman, 2003) or criteria derived from information theory (Bekkerman et al., 2003; Forman, 2003; Torkkola, 2003). A particular disadvantage of the ranking approach is that it does not consider any dependencies amongst the features and so may select suboptimal features in the sense of predictive capability as a group.

Various group-selection approaches have been proposed to deal with that problem. The first category uses machine learning techniques (decision trees, FOCUS algorithm (Almuallim and Dietterich, 1991, 1994), Relief algorithm (Kira and Rendell, 1992a,b; Kononenko, 1994), Naive-Bayes induction algorithms (Duda and Hart, 1973; Good, 1965; Langley et al., 1992)) to evaluate the predictive power of a subset of features and then chooses the subset with the best predictive capability (Kohavi and John, 1997). An obvious shortcoming of this approach is that the feature set selected depends on which particular machine learning technique is used. Another difficulty is related to the large search space, i.e., the power set of features whose size exponentially increases. Strategies that have been proposed to deal with the search efficiency include best-first, branch-and-bound, simulated annealing, randomized hill climbing and genetic algorithms (see (Kohavi and John, 1997) for a review). Each strategy has pros and cons, and a discussion on this can be found in (Guyon and Elisseeff, 2003; Saeys et al., 2007).

The second category of group-selection approaches searches for the optimal feature set by choosing the best feature one-by-one according to some intermediate criterion such as Euclidean distance, information gain or gain ratio (Ben-Bassat, 1982), or Chi-squared coefficient. Multivariate feature selection such as correlation-based feature selection (Hall, 1999), Markov blanket filter (Koller and Sahami, 1996) and fast correlation-based feature selection (Yu and Liu, 2004) has been reported as well. The incremental tree we use in this paper belongs to this category and the intermediate criteria we chose are correlation and cross entropy.

The last category of group-selection approaches performs feature selection together with model training. This means that the selection approach is specific to the model that is being trained. This approach has been studied in the context of decision trees such as CART (Breiman et al., 1984) and random forests (Liaw and Wiener, 2002), weighted naive Bayes (Duda et al., 2001), weight vector of SVMs (Guyon et al., 2002; Weston et al., 2003) and

regression/selection by Lasso (Tibshirani, 2011).

It is also possible to combine different selection approaches (Bi et al., 2003). Interested readers are referred to (Guyon and Elisseeff, 2003; Saeys et al., 2007) for more details regarding various selection methods and their combination.

2.1.3. Feature combination

Although many sorts of models can be used to combine multiple features, discriminative models have an advantage when the decision boundary is complex because they do not assume any form of distribution over the data. This means they are well-suited for combining features derived from heterogeneous sources, which may have different dynamic ranges or even be represented in different ways (e.g., continuous features and discrete features). We therefore focus our review on discriminative model-based feature combination and confidence estimation.

Among various discriminative models, MLPs (Mathan and Miclet, 1991; Weintraub et al., 1997; Ou et al., 2001; Vergyri et al., 2006), linear discriminative functions (Sukkar and Wilpon, 1993; Gillick et al., 1997; Kamppari and Hazen, 2000), GLMs and GAMs (Siu et al., 1997), DTs (Neti et al., 1997; Hauptmann et al., 1998) and SVMs (Zhang and Rudnicky, 2001; Ayed et al., 2002; Sudoh et al., 2006; Shafran et al., 2006) are the most popular. Comparisons between DTs, GLMs, GAMs and neural networks have been carried out by Chase (1997), and linear classifiers and neural networks are compared in (Schaaf and Kemp, 1997). Both studies concluded that neural networks outperformed the other discriminative models. Hernández (2000) compared linear discriminant functions, neural networks and fuzzy logic-based systems, and reported that the fuzzy logic-based systems provided the best result. Zhang and Rudnicky (2001) studied DTs, neural networks and SVMs, and reported that the best performance was given by the SVM. Wang et al. (2012) employed MLPs, SVMs and DTs and concluded that all discriminative models provided marginal performance gains and that which model is superior is determined by the properties of the query terms, e.g., whether they are in-vocabulary or out-of-vocabulary.

2.2. Discriminative confidence normalization

We have mentioned that discriminative models can be employed to combine heterogeneous features and produce a discriminative confidence measure for decision making. This confidence measure, however, may not be directly

suitable for STD since it is not optimal with respect to the evaluation metric, i.e., ATWV. In addition, we wish to employ more term-dependent features in confidence estimation so that the diversity among search terms can be compensated for. This is particularly important for OOV term detection, in the light of high diversity among OOV terms. The discriminative confidence normalization framework, which is motivated by the term-specific threshold (TST) approach proposed by Vergyri et al. (2007), aims to address these concerns. Since all our STD experiments are conducted based on this framework, we now present a brief overview of it.

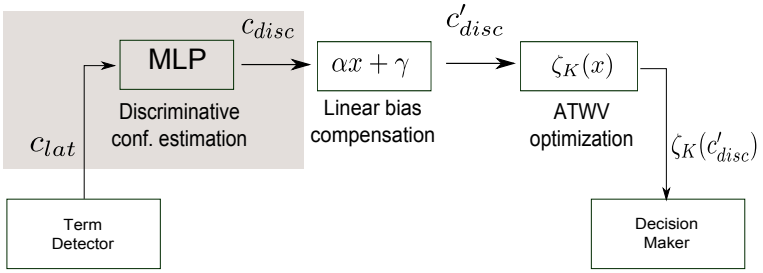

Figure 2: The diagram of discriminative confidence estimation and normalization for STD.

Figure 2 illustrates the framework, where multiple features including the lattice-based confidence c_{lat} are fed into an MLP model, from which the discriminative confidence c_{disc} is generated. A linear transform is then applied to compensate for possible bias introduced by lattice approximation and discriminative training. Finally, the de-biased discriminative confidence c'_{disc} is normalized by a transform ζ_K to compensate for the diverse occurrence rate of different terms. The resulting confidence $\zeta_K(c'_{disc})$ is unbiased, discriminative, term-dependent and optimal for the STD metric ATWV. Building such a framework requires training the MLP and estimating the parameters α and γ for the linear transform; more details can be found in (Wang et al., 2012).

The work in this paper focuses on the first part of this framework, i.e., discriminative confidence estimation (shaded in Figure 2). Our goal is to find a set of features that contribute most to STD, and then use them in the framework of Figure 2 to produce a discriminative confidence measure that is suitable for STD. Therefore, we need to build the entire STD architecture and the discriminative confidence framework to test the analysis results within a

practical STD system.

2.2.1. MLP-based discriminative confidence

An MLP has been used to derive the discriminative confidence. This MLP is a feed-forward artificial neural network that maps input features into an output response. The structure of the MLP is comprised of an input layer, which contains as many units as the number of input features, a hidden layer with a sigmoid activation, and an output layer with a softmax activation, as shown in Figure 3. This output layer contains two units, corresponding to the hit and FA classes respectively. The Weka tool (Hall et al., 2009) has been used to build the MLP structure and to derive the discriminative confidence in Figure 2.

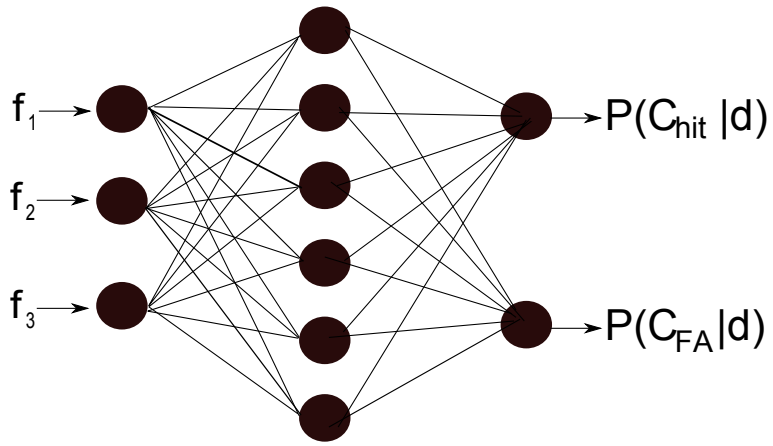

Figure 3: The MLP structure used to derive the discriminative confidence with three input features f_1, f_2 and f_3.

3. Experimental setup

In order to set the analysis presented in following sections in context, we first introduce the experimental settings used in this work. This involves the speech, text and lexicon data we used for model training, system development and performance evaluation, as well as the configurations of our STD experiments and feature analysis.

3.1. Data profile

The data profile involves search terms, speech and text databases for training the models, tuning parameters and evaluating performance. We selected two languages on which to conduct our study: English (which is the most common language used in STD research) and Spanish. By comparative study, we hope to identify universal features that are generally effective for multiple languages/domains, plus some language/domain-dependent features that are important for particular languages/domains. We present the data profile for each language in turn below.

3.1.1. English data

For English, our STD system operates in the domain of multi-party meetings. We first chose 557 single-word terms from the 50k dictionary used by the AMI RT05s LVCSR system (Hain et al., 2006) for the purpose of system development, among which 490 terms were used to train the MLP model for discriminative confidence estimation and 67 terms were used to tune the rest of the parameters, particularly the parameters in linear bias compensation (see Figure 2). In addition, a disjoint set of 484 single and multi-word terms was chosen for performance evaluation. All these terms are OOV - we made this choice because OOV terms are much more difficult to detect and score compared to INV terms, and tend to gain more from the discriminative confidence normalization framework which is the subject of our study. Figure 4 shows the length distribution of the evaluation terms.

The speech data are recorded with individual headset microphones at several sites, including ICSI, NIST, ISL, LDC, the Virginia Polytechnic Institute and State University and various partners of the AMI project. The training set, which is used to train the acoustic models (AMs) of the ASR subsystem, is the same as the one used to train the AMI05 LVCSR system (Hain et al., 2006), except that all utterances containing the evaluation terms were deleted. This data purging ensures the evaluation terms are truly OOV in every respect, including acoustic model training. After OOV purging, 80.2 hours of speech were left to train the AMs. The development set, which is used to tune system parameters and collect training samples for the MLP model in discriminative confidence estimation, is the standard NIST RT04s dev set. Finally, the evaluation set consists of the NIST RT04s and RT05s eval sets plus the speech corpus AMI08 recorded at the University of Ed-

inburgh in the AMIDA project[1]. The total size of the evaluation set is 11 hours of speech which contains 2735 occurrences of the evaluation terms.

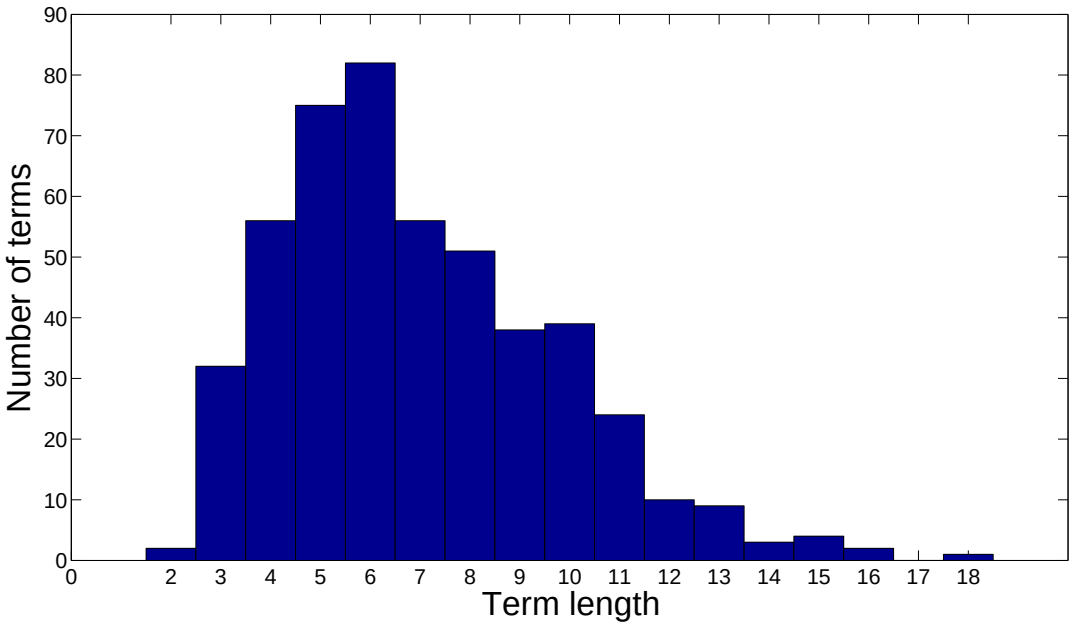

Figure 4: *Term length distribution of the English data on the evaluation set.*

The text corpus that was used to train the language models for the ASR subsystem was provided by the AMI project and is the same as the one used by the AMI RT05s LVCSR system (Hain et al., 2006). This corpus contains text from various sources such as news, transcripts of speech corpora and a large amount of web text, amounting to 521.4M words. The 50k AMI dictionary was used to convert the word-based text corpus to a phoneme-based corpus.

Feature analysis can be regarded as a step in system development. The same speech data and search terms used to train the MLP model are employed to train the analysis models (linear regression and logistic linear regression) and perform feature quality check.

[1] *http://www.amiproject.org*

3.1.2. Spanish data

For Spanish, the Albayzin database (Moreno et al., 1993) was selected to build the system and conduct the analysis. This database involves reading style speech recorded in a silent environment. Some of the sentences are chosen from the general domain and others are from the geographical domain. The geographical domain utterances involve entity names in Spain, such as mountains, rivers and cities. Phonetic balance was considered in the database design. The training data involve 3.3 hours of speech from the general domain, and the development and evaluation sets involve 3.6 hours and 2 hours of speech in the geographical domain, respectively. Finally, the transcripts of the training speech data are used to train language models; due to the simple grammar of the utterances, this naive approach works well in phone recognition.

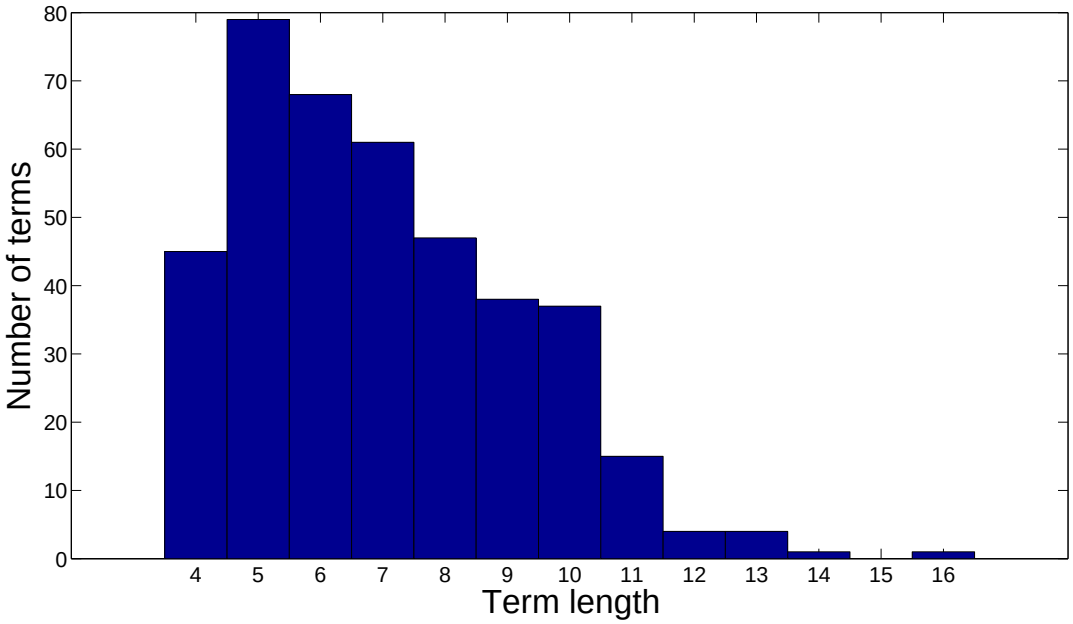

Figure 5: *Term length distribution of the Spanish data on the evaluation set.*

For STD, we chose 605 single-word terms from the development utterances to conduct system development, among which 500 terms were used to train the MLP model for discriminative confidence estimation and 105 terms were used to tune other parameters. Finally, a disjoint set of 400 single-word

terms including name entities (mountains, rivers, cities, etc) and common words from the evaluation utterances was selected to evaluate STD. These terms, which are again all OOV, occur a total of 11329 times in the evaluation data. As in English experiments, feature analysis employs the same data and search terms used to construct the MLP, to train the analysis models and evaluate feature quality. Figure 5 shows the length distribution of the evaluation terms.

Comparing the two languages, we see that the English data are more conversational and spontaneous, while the Spanish data are more clean and constrained. This contrast allows a comparative study. We note that a search term may involve one or several words. In phone-based STD, however, the single-word and multi-word terms are treated in a unified way since all the terms are converted into phone sequences. For a clear presentation of the data profile, Table 1 shows the top-20 terms that occur most frequently in the English and Spanish evaluation data.

3.2. System configuration

For both English and Spanish experiments, we built phoneme-based STD systems. The acoustic models are state-clustered triphones built on 39 dimensional Mel frequency cepstral coefficient (MFCC) features using the HTK tool (Young et al., 2006). The n-gram language models (LM) were built using the SRI LM toolkit. For English, we choose a 6-gram model due to its high performance on the development set; for Spanish, we choose a simpler 2-gram due to the limited amount of training text and the simple grammar of the utterances.

For STD, the *Lattice2Multigram* tool developed by the Speech@FIT group of the Brno University of Technology (BUT) is used to conduct lattice search and retrieve potential term occurrences. We have extended this tool to support complex confidence measures, including the discriminative confidence studied in this work. Additionally, certain letter-to-sound (LTS) approaches need to be designed for inferring pronunciations of OOV terms. For English, we employed an enhanced joint-multigram model (Deligne et al., 1995; Wang et al., 2009a) trained with the AMI dictionary for this purpose; for Spanish, we chose a simple mapping approach to derive the pronunciation. More detailed information about the experimental settings can be found in (Wang, 2009) and (Tejedor, 2009).

	English			Spanish		
ID	Term	#Occ.	# Phones	Term	#Occ.	# Phones
1	remote	310	5	comunidad	536	9
2	control	216	7	dime	504	4
3	minutes	73	6	ríos	474	4
4	project	65	7	nombre	300	6
5	information	57	8	comunidades	248	11
6	marketing	57	8	pasa	244	4
7	budget	55	5	autónoma	232	8
8	speech	52	4	ciudades	222	8
9	target	49	6	mayor	200	5
10	recognition	48	9	cuál	196	4
11	euros	46	5	habitantes	196	9
12	fifty	43	5	caudal	190	6
13	television	41	8	pico	166	4
14	course	41	4	dónde	166	5
15	features	36	5	picos	162	5
16	twenty	34	6	metros	160	6
17	question	31	7	tiene	158	5
18	system	27	6	sistema	150	7
19	email	26	4	mediterráneo	148	11
20	mouse	26	3	longitud	144	8

Table 1: *Top* 20 *most frequent terms of the evaluation set, their number of occurrences and their number of phones for English and Spanish data.*

4. Features in analysis

This section presents the features studied in this work. Some of the features are motivated by the relevance analysis on ASR in (Goldwater et al., 2009) and others were selected due to properties of STD. Table 2 summarizes the features, where each feature is assigned a number to assist the presentation in the following sections.

These features are categorized into six groups, according to the information source: lattice-based features, lexical features, Levenshtein distance-based features, duration-based features, position-based features and prosodic features. The prosodic features can be further categorized into two subgroups: pitch-based and energy-based.

- Lattice-based features: This group of features includes the lattice-based confidence $c_f(d)$, the effective occurrence rate $R_0(K)$, and the effective false alarm rate $R_1(K)$ defined as follows:

$$R_0(K) = \frac{\sum_i c_f(d_i^K)}{T} \tag{3}$$

and

$$R_1(K) = \frac{\sum_i (1 - c_f(d_i^K))}{T} \tag{4}$$

where $c_f(d_i^K)$ represents the lattice-based confidence of the i-detection of the term K and T is the total length of the audio in seconds.

Other features belonging to this group include the maximum (max), minimum (min) and mean acoustic scores and language model scores of all the phones of the detected term. In general, the lattice-based features reflect reliability of the putative detection under consideration, as compared to alternative terms in the same speech segment and to the acoustic and language model score distribution over phones of terms.

- Lexical features: This group of features includes the total number of graphemes, phones, vowel graphemes, consonant graphemes, vowel phones and consonant phones for each term. These features reflect lexical properties of search terms, particularly term length and crude information about the consonant-vowel structure.

Feature number	Feature	Type
1	Lattice-based confidence ($c_f(d_i^K)$)	Lattice-based
2	Effective occurrence rate ($R_0(K)$)	Lattice-based
3	Effective FA rate ($R_1(K)$)	Lattice-based
4	Max phone acoustic score	Lattice-based
5	Min phone acoustic score	Lattice-based
6	Mean phone acoustic score	Lattice-based
7	Max phone language model score	Lattice-based
8	Min phone language model score	Lattice-based
9	Mean phone language model score	Lattice-based
10	Number of graphemes	Lexical
11	Number of phones	Lexical
12	Number of vowel graphemes	Lexical
13	Number of consonant graphemes	Lexical
14	Number of vowel phones	Lexical
15	Number of consonant phones	Lexical
16	Max Levenshtein distance	Levenshtein distance
17	Min Levenshtein distance	Levenshtein distance
18	Mean Levenshtein distance	Levenshtein distance
19	Duration	Duration
20	Phonetic speech rate	Duration
21	Vowel speech rate	Duration
22	Max phone duration	Duration
23	Min phone duration	Duration
24	Mean phone duration	Duration
25	Position	Position
26	Max pitch	Prosodic (pitch)
27	Min pitch	Prosodic (pitch)
28	Mean pitch	Prosodic (pitch)
29	Max phone pitch	Prosodic (pitch)
30	Mean phone pitch	Prosodic (pitch)
31	Min phone pitch	Prosodic (pitch)
32	Voicing percentage	Prosodic
33	Max energy	Prosodic (energy)
34	Min energy	Prosodic (energy)
35	Mean energy	Prosodic (energy)
36	Max phone energy	Prosodic (energy)
37	Mean phone energy	Prosodic (energy)
38	Min phone energy	Prosodic (energy)

Table 2: *Features and types of features under analysis.*

- Levenshtein distance features: This group includes the maximum, minimum and mean Levenshtein distance from the current term to the other terms in the vocabulary. These features try to measure the degree of confusability, and therefore are highly dependent on the vocabulary used.

- Duration features: This group of features consists of the duration of the detection, the 'phonetic speech rate' (term duration divided by the number of phones of the term) and the 'vowel speech rate' (term duration divided by the number of vowels of the term). In addition, the maximum, minimum and mean phone duration also belong to this group. These features provide information in two ways: the speech rate is directly related to reliability of the ASR output; the distribution of the duration over phones can be useful in identifying abnormal detections.

- Position features: This group has just a single feature which represents the position of the detection. There are three values that this feature can take: the beginning of the lattice, the end of the lattice, or another position. This feature is included to represent the fact that speech decoders have less reliability at the beginning and end of utterances.

- Prosodic features: This group of features reflects the prosody of speech: pitch (maximum, minimum and mean pitch for each detection), energy (maximum, minimum and mean energy for each detection) and voicing percentage (i.e., the percentage of voiced speech for each detection in the speech signal). As additional features, they also include the maximum and minimum pitch and energy values of all the phones corresponding to the term detected and the mean pitch and energy per phone. All these features were computed using Praat (Boersma and Weenink, 2007). Together, they provide some information about pitch and energy dispersion and abnormal pitch or energy values.

To summarize: this work extends our previous study (Tejedor et al., 2010) by including a number of new features, particularly the phone-based features (i.e., acoustic and language model scores, phone duration, and pitch and energy per phone). Compared with the studies reported by others, such as Ou et al. (2001); Ayed et al. (2002); Wang et al. (2012), the novel features in this work are those involving: Levenshtein distances, position, prosodic

features, all the lexical features except the number of phones, and all the duration features except the duration of the detection.

5. Individual feature analysis

In this section, we investigate the contribution of the features presented in the previous section. The goal of the investigation is to select the most informative according to their contribution to STD performance. As discussed already, conducting this analysis using the MLP in conjunction with the STD metric ATWV is possible but computationally costly and risks losing generalization capacity. We therefore use simpler models to obtain rankings according to intermediate evaluation metrics. In this work we use linear regression and logistic linear regression as the analysis tools, and correspondingly, choose the regression residual (R^2) and cross entropy as the evaluation metrics for LR and LLR respectively. Generally speaking, LR looks for features that are most *relevant* to the detection decision, while LLR targets the features that are most *important* for the decision and therefore is probably more directly related to the STD results.

The feature analysis starts with data preparation. First of all, STD experiments were conducted on the development set and a set of putative detections was obtained. Each detection was assigned a decision variable, which is 1 for hits and 0 for FAs. Combining the features presented in the previous section and the decision variable, these detections form the training dataset for the analysis models. Given that the number of FAs is much higher than the number of hits, this training set was first balanced in terms of hits and FAs by adding randomly selected repetitions of hits until there was an equal number of hits and FAs. In order to avoid over-fitting, we divided this set into two equal-size subsets and used the first part for training the models (learning set) and the second part for evaluation (verification set) in terms of R^2 and/or cross entropy.

5.1. Histogram-based analysis

We first present an intuitive analysis of the discriminative power of each individual feature. First, we collect all detection instances from the development set belonging to the hit and FA classes, and then for each feature, we draw distributions of the feature values within the two classes. The overlap proportion of the two distributions reflects the discriminative capability of the feature. The results are shown in Figures 6–17. We can observe that

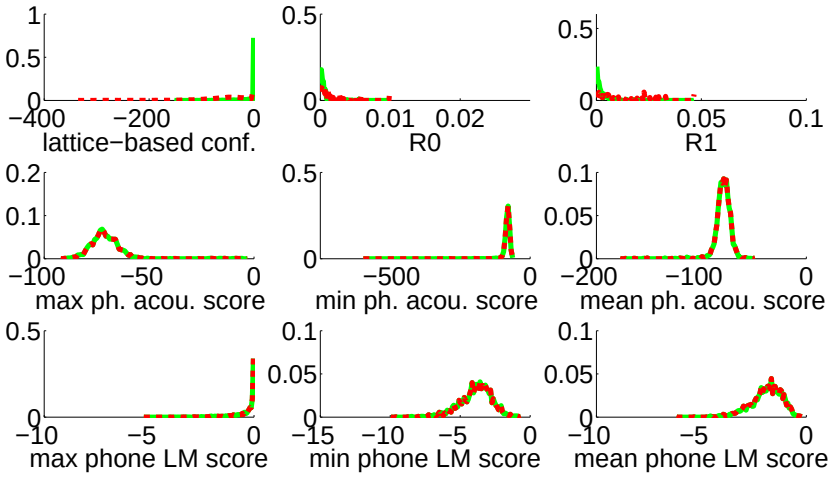

Figure 6: *Distributions of the lattice-based features for English data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs. 'ph.' denotes phone, 'acou.' denotes acoustic and 'conf.' denotes confidence.*

the lattice-based features and the lexical features posses good discriminative capability in general. The position feature and the prosodic features (pitch, voicing percentage) exhibit less discriminative power. The energy-based features do not contribute much either, except the min energy in Spanish data that exhibits moderate discrimination. The contribution of the Levenshtein distance-based features is also moderate. In addition, some features behave differently in the English and the Spanish experiments. For instance, the duration-based features contribute less significantly to the English system than to the Spanish system. This may be attributed to the unreliable duration estimation with the English data that belong to the meeting domain, and therefore are highly spontaneous. More detailed analysis will be given in the next section.

5.2. Regression analysis

In this section, we employ linear regression to study the relevance of features to the decision results, i.e., hits/FAs. LR can be simply formulated as follows:

$$y = WF \tag{5}$$

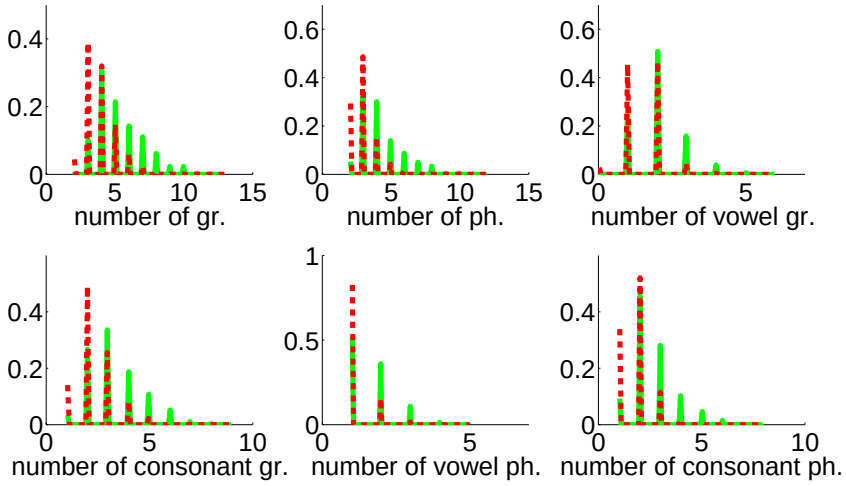

Figure 7: *Distributions of the lexical features for English data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs. 'ph.' denotes phones and 'gr.' denotes graphemes.*

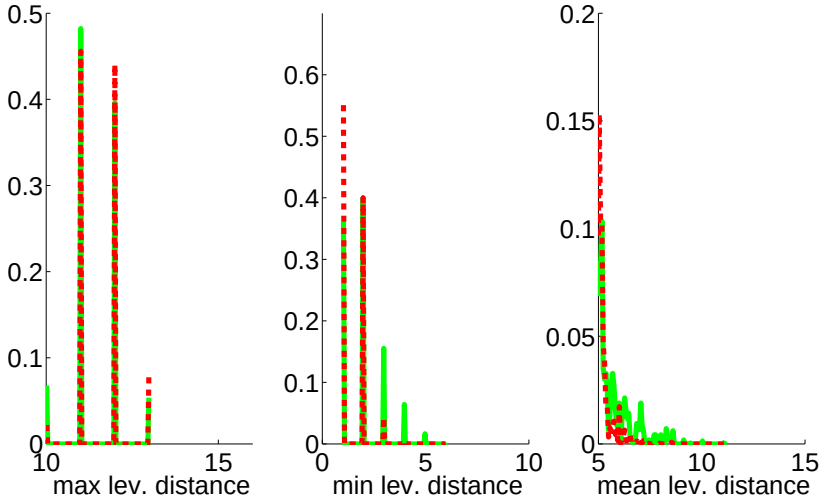

Figure 8: *Distributions of the Levenshtein distance-based features for English data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs. 'lev.' stands for Levenshtein.*

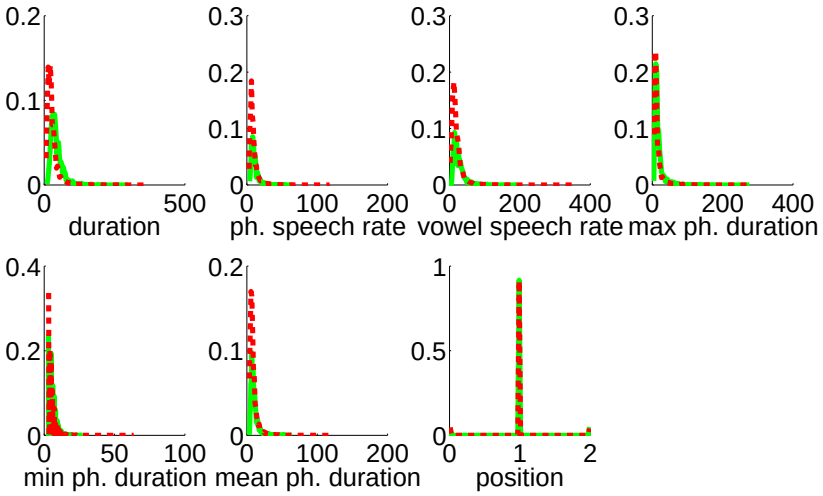

Figure 9: *Distributions of the duration-based features and the position feature for English data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

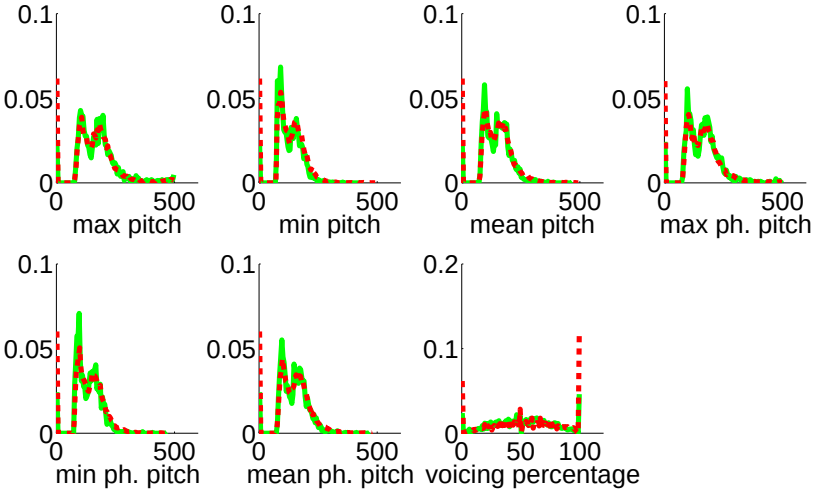

Figure 10: *Distributions of the pitch-based features and the voicing percentage feature for English data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

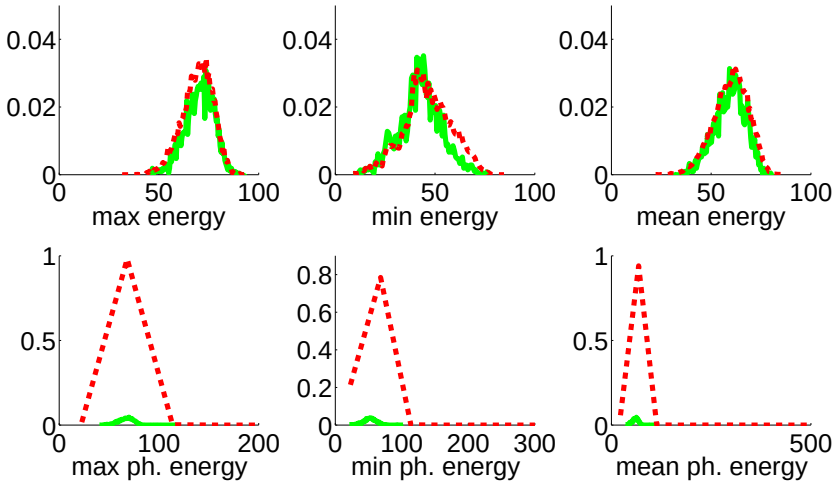

Figure 11: *Distributions of the energy-based features for English data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

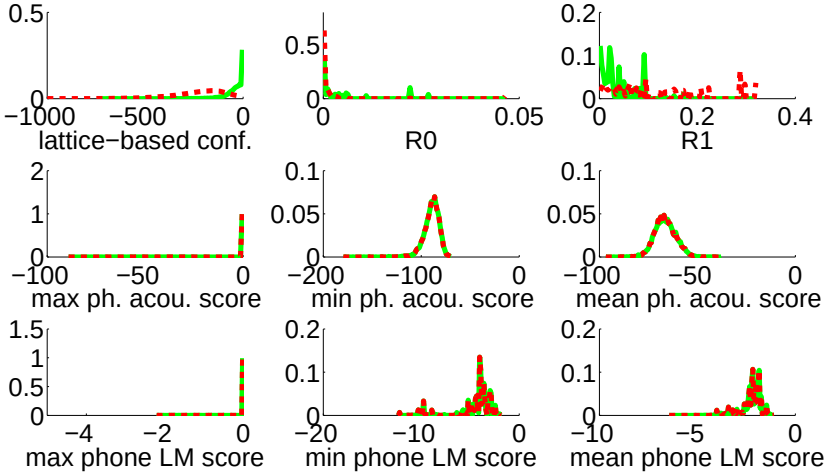

Figure 12: *Distributions of the lattice-based features for Spanish data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

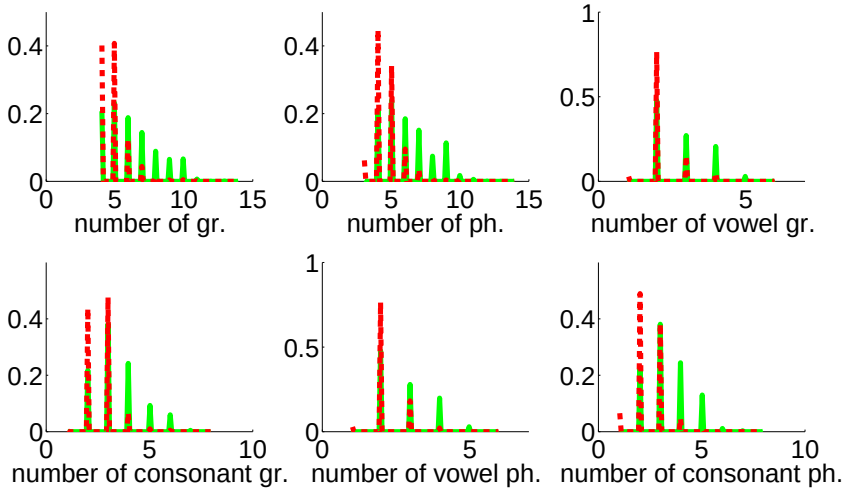

Figure 13: *Distributions of the lexical features for Spanish data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

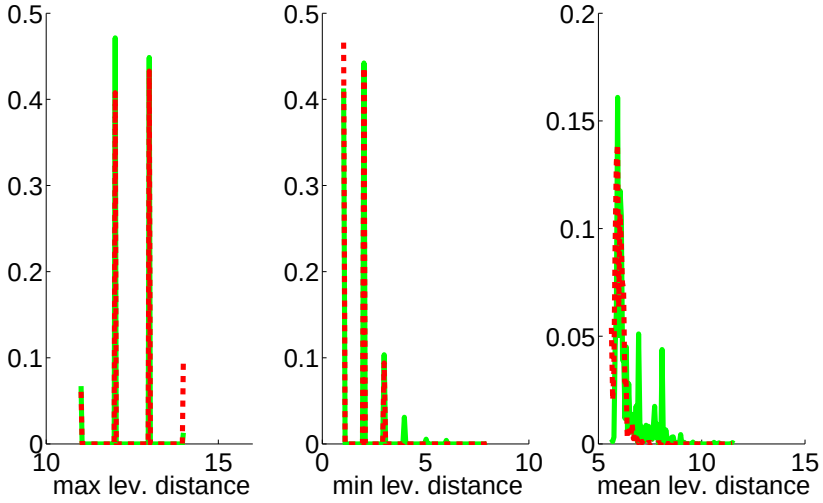

Figure 14: *Distributions of the Levenshtein distance-based features for Spanish data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

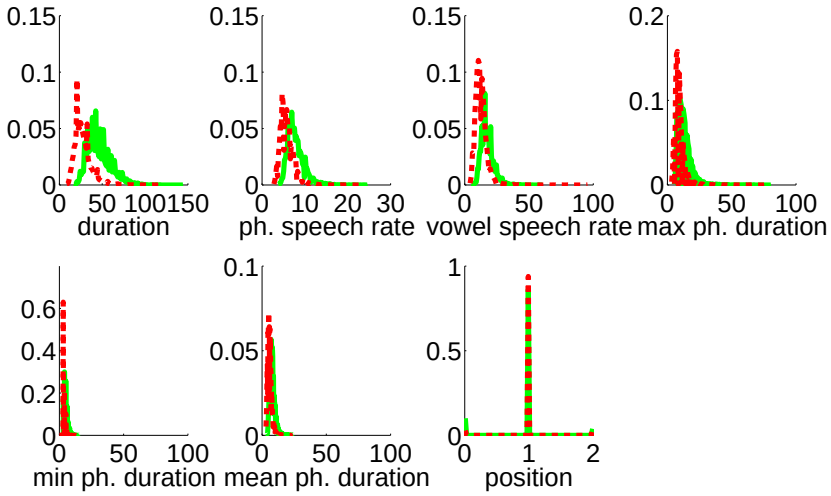

Figure 15: *Distributions of the duration-based features and the position feature for Spanish data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

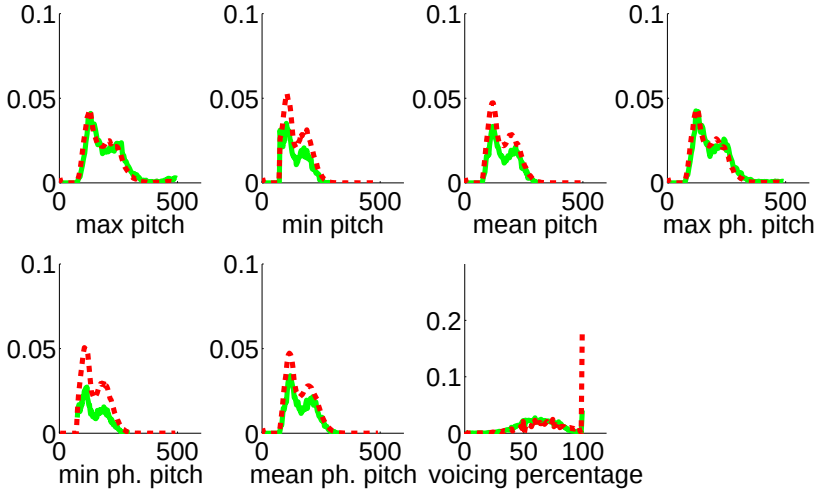

Figure 16: *Distributions of the pitch-based features and the voicing percentage feature for Spanish data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

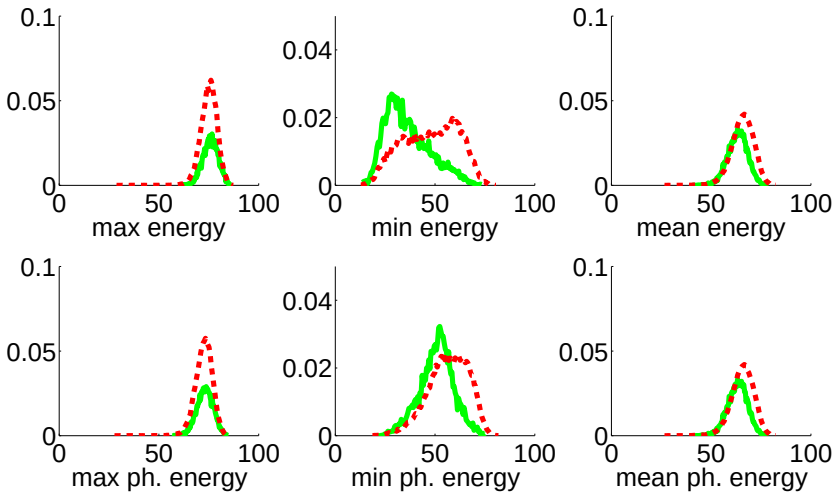

Figure 17: *Distributions of the energy-based features for Spanish data on the whole development set. The solid green (light grey) curve represents hits and the dashed red (dark) curve represents FAs.*

where F is a $(R+1) \times 1$ matrix involving the R features participating in the regression plus a constant element for bias, and W holds the parameters of the model in a $1 \times (R+1)$ matrix – these are the weights on the features in F. W is usually trained with respect to squared error defined as follows:

$$E = \frac{1}{N} \sum_{i=1}^{N} (y_i - t_i)^2 \tag{6}$$

where N is the number of detections in the learning/verification set, and y_i is the output of the regression model for the i-th detection. t_i is the decision label, which has been coded as 1 and 0 for hits and FAs respectively. Note that the squared error E is directly related to the Pearson correlation coefficient R^2, defined as follows:

$$R^2 = \frac{S_{ty}^2}{S_{tt} S_{yy}} \tag{7}$$

where S_{ty} is the covariance of t and y and is defined as

$$S_{ty} = \frac{1}{N}\sum_{i=1}^{N}(t_i - \bar{t})(y_i - \bar{y}) \tag{8}$$

where $\bar{t}$ and $\bar{y}$ represent the mean of t and y respectively. S_{tt} and S_{yy} are obtained using the same formula by replacing t with y or vice versa. The squared error E and R^2 are directly related:

$$R^2 = 1 - \frac{\tilde{E}}{S_{tt}} \tag{9}$$

where $\tilde{E}$ is the squared error obtained with the optimal parameter W. For this reason we use R^2 as the evaluation metric in LR analysis.

As a starting point, we use LR to analyze individual features. In this configuration, the resulting R^2 represents the relevance of each feature to the decision result. Figure 18 presents R^2 obtained in the verification set with all individual features for the two languages under investigation. We can see that the lattice-based confidence (#1) is the most relevant for both languages. This is not surprising since this feature incorporates a multitude of useful information derived from ASR decoding and represents a theoretically-sound confidence measure. We also find that there are quite a few features besides the lattice-based confidence that exhibit high relevance to the decision, e.g., effective FA rate (#3), lexical features (#10 - #15) and Levenshtein distance-based features (#17 and #18). The effective FA rate is prominent because it is designed to approximate the FA prior, and the lexical and Levenshtein distance-based features are directly related to the confusion degree of a term and thus convey information about potential errors. Finally, we find features derived from some groups such as pitch are generally less relevant, indicating that the pitch information, which may be more relevant to long acoustic contexts, is less informative for the STD task which focuses on local detections. This confirms the findings of the histogram-based analysis reported previously.

An interesting observation is that the features derived from duration and energy exhibit much more significant relevance to decision in Spanish than in English data. This has been observed in the histogram analysis. The different speaking styles of the English data (spontaneous meeting data) and Spanish data (clean read speech data) may contribute to the difference. First of all, we notice that duration is a good prior for error detection. For example, ASR tends to produce more errors for short words (Goldwater et al.,

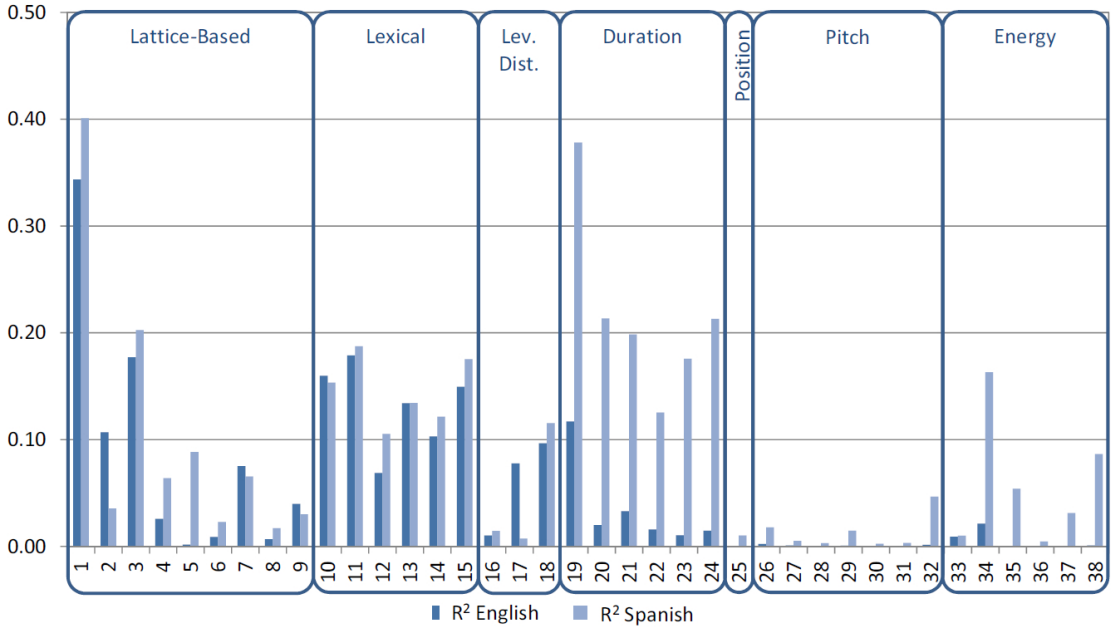

Figure 18: R^2 *obtained for each individual feature for English and Spanish data on the verification set. "Lev. Dist." refers to the Levenshtein distance-based features.*

2009; Kao et al., 2011) and compensating for duration reduces the word error rate (Gadde, 2000). In STD, decision errors tend to be reduced by integrating duration-based features, if they are reliable. This explains why duration contributes to the Spanish data. For the English data which are spontaneous meeting speech, however, the speaking style varies in a significant way, which leads to unreliable duration-based decision. The same argument applies to the energy-based features. In addition, ASR is inherently more difficult for meeting data than read speech, so more noise can be expected in features relating to duration (remembering that these are obtained from an alignment between the acoustic models and the speech data during decoding); consequently, noise is propagated into all other features which somehow rely on this alignment (e.g., values of pitch or energy for specific phones). This noise leads to less discriminative features for correct/incorrect term detections. In contrast, the Spanish data are rather clean and constrained in speaking style, and these features are more informative; for example, abnormal duration and energy are good indicators for potential errors.

To enable a clearer understanding, we limit ourselves now to discussion of

	English		Spanish	
Rank	Feature	R^2	Feature	R^2
1	Lattice-based confidence	0.3435	Lattice-based confidence	0.4010
2	Number of phones	0.1787	Duration	0.3781
3	Effective FA rate	0.1770	Phonetic speech rate	0.2135
4	Number of graphemes	0.1597	Mean phone duration	0.2131
5	Number of consonant phones	0.1496	Effective FA rate	0.2026
6	Number of consonant graphemes	0.1341	Vowel speech rate	0.1986
7	Duration	0.1169	Number of phones	0.1873
8	Effective occurrence rate	0.1070	Min phone duration	0.1758
9	Number of vowel phones	0.1030	Number of consonant phones	0.1755
10	Mean Levenshtein distance	0.0967	Min energy	0.1632

Table 3: *Top* 10 *relevant features obtained with LR for English and Spanish data.* R^2 *results on verification set are also reported for both languages.*

the top 10 relevant features, as presented in Table 3 for English and Spanish data. We observe that 5 features out of the top 10 are shared across English and Spanish data: lattice-based confidence (which is always the most relevant feature), effective FA rate, duration, number of phones and number of consonant phones. This agreement suggests that these features are generally informative for confidence estimation and decision making in STD. Again, we find that most of the top 10 features are related to lattice statistics and lexical properties for English data, while for Spanish data, a large portion of the top 10 features are derived from lattice statistics, duration and energy.

5.3. Discriminative analysis

In the discriminative analysis, we use logistic linear regression to test the capability of each individual feature to distinguish hits from false alarms. LLR is a simple extension of linear regression by compressing the regression result to the range $[0, 1]$ using the logistic function, and thus belongs to the family of generalized linear models. It is formulated as follows:

$$y = \sigma(WF) \tag{10}$$

where W and F are parameters and features respectively, and $\sigma(x)$ is the sigmoid function given by:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{11}$$

Theoretically, LLR outputs the posterior probabilities of detections being hits given the features, assuming a wide range of distributions over those features for the hit and FA classes. The objective of LLR is cross entropy (CE), defined as follows:

$$CE = -\frac{1}{N}\sum_{i=1}^{N} t_i log y_i - \frac{1}{N}\sum_{i=1}^{N}(1 - t_i)log(1 - y_i) \tag{12}$$

where N is the number of detections in the learning/verification set and y_i and t_i are the output of LLR and the decision label of i-th detection respectively. In contrast to R^2, cross entropy has a probabilistic interpretation and is more related to discriminative behavior, and therefore LLR analyzes the contribution of features to detection results from the perspective of discriminative power instead of relevance.

In this section we use LLR to analyze individual features, with cross entropy as the metric. Figure 19 presents the results for English and Spanish data in the verification set, and Table 4 presents the top 10 discriminative features. Note that, in both representations, lower CE indicates better discrimination. We first observe that both for English and Spanish data, features derived from lattices, lexical properties, and Levenshtein distance characteristics are among the most informative, and pitch features are unimportant. Features derived from duration and energy exhibit much more importance for Spanish than for English data. All these observations are consistent with those we obtained with the LR analysis. The consistency is maintained when looking at the top-10 most discriminative features: LR and LLR result in very similar top 10 lists. This consistency indicates that the most discriminative features for decision are also those most relevant to the decision results.

5.4. Leave-one-out feature analysis

The leave-one-out analysis is another way to investigate contribution of individual features. Although closely related to the individual analysis, the leave-one-out experiment examines how much is lost if a feature is removed from the feature set, and therefore reflects how much *unique* and *additional* contributions a particular feature provides.

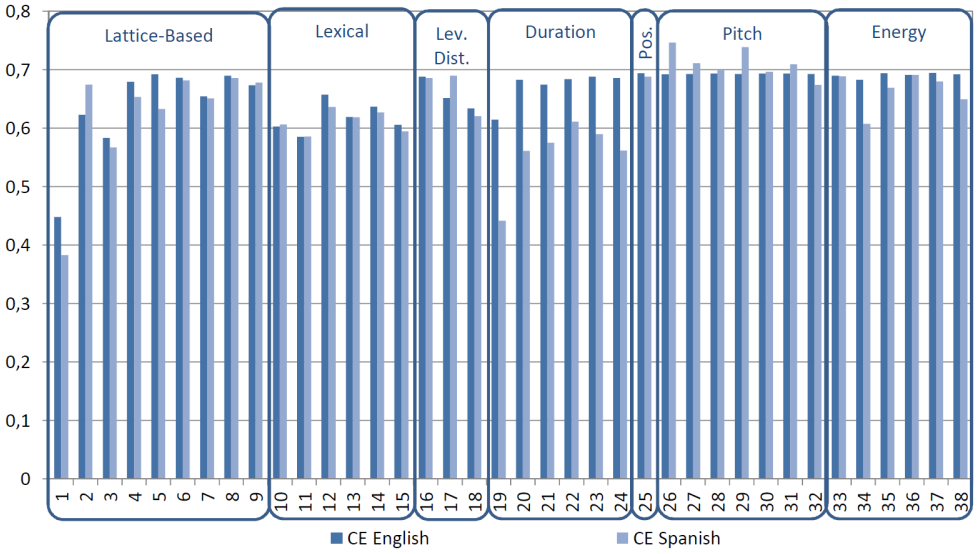

Figure 19: *CE obtained for each individual feature for English and Spanish data on the verification set. "Lev. Dist." refers to the Levenshtein distance-based features. "Pos." stands for the Position feature.*

	English		Spanish	
Rank	Feature	CE	Feature	CE
1	Lattice-based confidence	0.4483	Lattice-based confidence	0.3828
2	Effective FA rate	0.5833	Duration	0.4414
3	Number of phones	0.5850	Phonetic speech rate	0.5611
4	Number of graphemes	0.6026	Mean phone duration	0.5614
5	Number of consonant phones	0.6057	Effective FA rate	0.5668
6	Duration	0.6147	Vowel speech rate	0.5751
7	Number of consonant graphemes	0.6189	Number of phones	0.5855
8	Effective occurrence rate	0.6225	Min phone duration	0.5899
9	Mean Levenshtein distance	0.6341	Number of consonant phones	0.5946
10	Number of vowel phones	0.6371	Number of graphemes	0.6064

Table 4: *Top* 10 *relevant features obtained with LLR for English and Spanish data. CE results on verification set are also reported for both languages.*

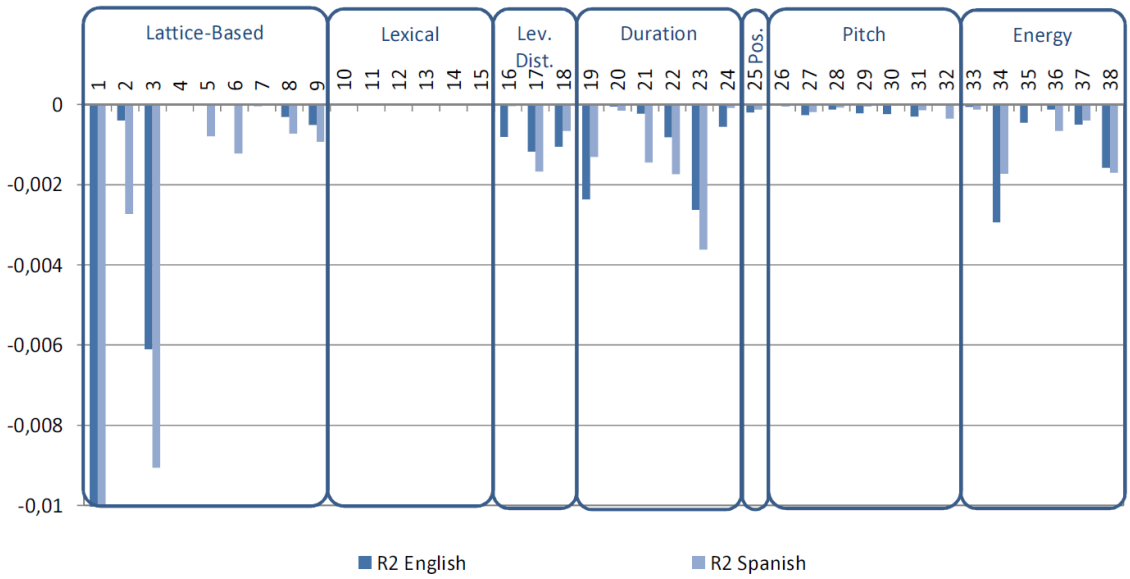

Figure 20: R^2 *reduction with LR in the leave-one-out experiment for English and Spanish data on the verification set. "Lev. Dist." refers to the Levenshtein distance-based features. "Pos." stands for the Position feature.*

Figure 20 and Figure 21 present the R^2 reduction in LR and the cross entropy increase in LLR for English and Spanish data in the verification set when a particular feature is removed from the feature set. We first observe that for both the English and Spanish data, missing the lattice-based confidence leads to the most significant R^2 reduction and cross entropy increase. Some other features exhibit similar but less significant change, such as the duration-based features (e.g., min phone duration), the Levenshtein distance-based features and the energy-based features. Note that these features are also in the top-10 most informative features in the individual analysis, indicating that they are not only informative by themselves but also complementary with others.

In contrast, some features can be removed without any change in R^2 and cross entropy, for instance, the lexical features. This does not mean that these features are less informative; instead, it is the high correlation with other features (e.g., duration) that marginalizes their contribution. This suggests that it is important to select features based on group contribution instead of individual performance, as we will present in the next section.

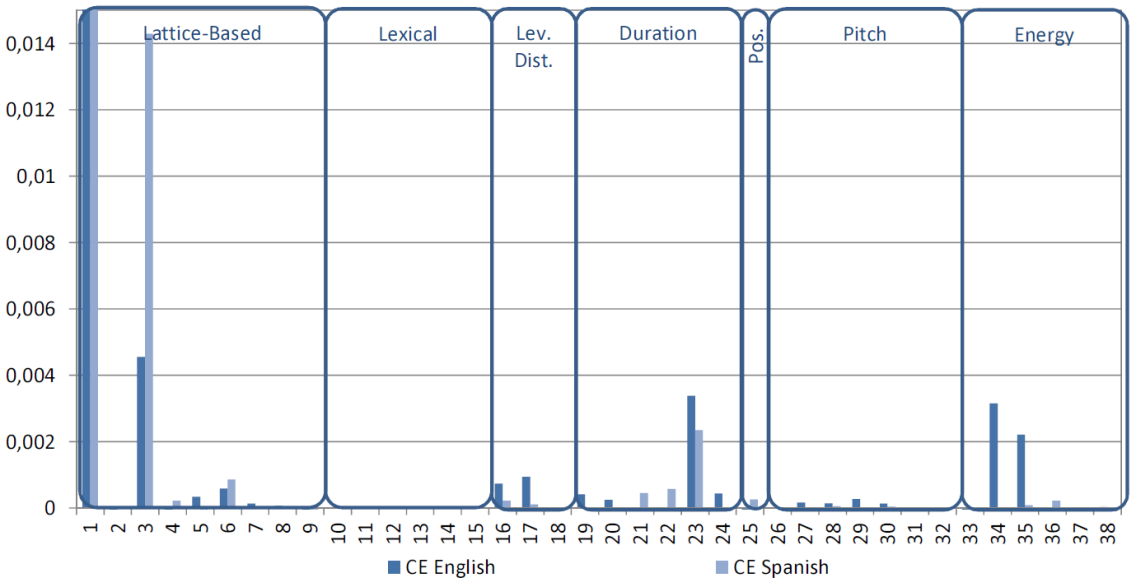

Figure 21: *Cross entropy increase with LLR in the leave-one-out experiment for English and Spanish data on the verification set. "Lev. Dist." refers to the Levenshtein distance-based features. "Pos." stands for the Position feature.*

6. Incremental feature analysis

In this section we select groups of features, noting that simply selecting the n individually-best features is suboptimal because of strong correlation among features. Figure 22 plots correlation matrices for English and Spanish data, where the grey level is proportional to the correlation between the pair of features indexed by the row and column numbers. We can see that some groups of features are strongly inter-related. In this work, we choose an incremental greedy selection approach, which enriches the n-best feature set one by one. At each step, the feature which provides the greatest increase in R^2 (for LR) or decrease in CE (for LLR) is selected.

6.1. Regression analysis

Again, we start with linear regression. The most relevant features are selected one by one by examining the R^2 increase caused by adding a new feature to the linear regression model. To investigate model generalization, we split the training samples into two equal-size subsets (as for the individual feature analysis), of which one is used for model training and feature selection (referred to as the *learning set*), and the other is used for evaluating

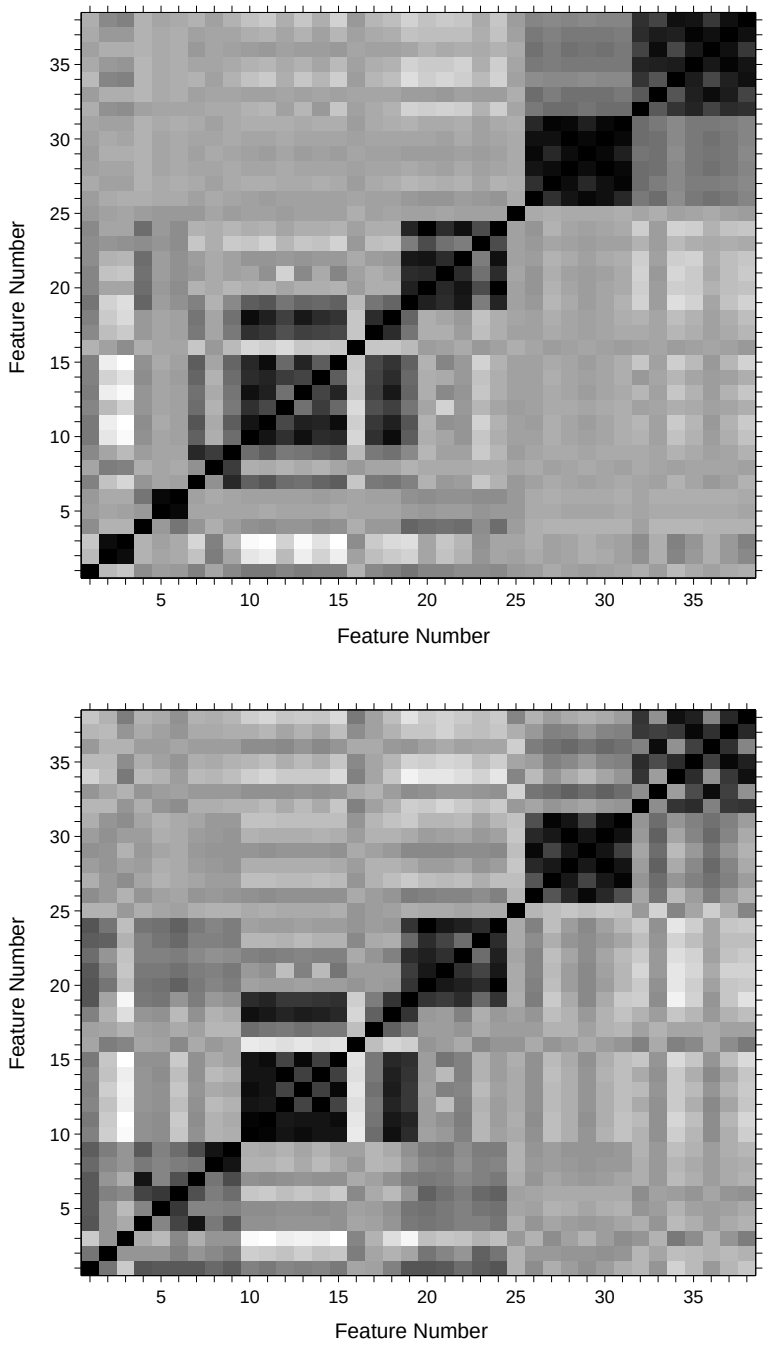

Figure 22: *The correlation matrices for English (top) and Spanish (bottom) data. Darker means stronger correlation (either positive or negative).*

the quality of the selection (referred to as the *verification set*). Since STD performance is highly related to hit/FA classification, we also present the Classification Error Rate besides R^2, which is defined as

$$CER = \frac{\sum_{i:y_i<0.5} t_i + \sum_{i:y_i \geq 0.5}(1 - t_i)}{N} \tag{13}$$

where y_i and t_i are regression output and decision label respectively, and N is the number of samples being tested.

Figure 23 presents the experimental results on English data, where the four curves represent respectively R^2 and CER obtained on the learning and verification sets. We observe that, on the learning set, the R^2 value increases rapidly as the first few features are added; with more features involved, only a little further increase in R^2 is obtained and the curve plateaus. This suggests that only the first $4-5$ features provide useful information and the remainder can be ignored. R^2 on the verification set displays a similar behavior: again, as the initial few informative features are added, R^2 increases quickly.

We also observe that the CER curves exhibit smooth descent, as features are added using LR-based incremental selection. This means that R^2 and CER are closely correlated, and that features selected based on R^2 are likely to be those with the best discriminative power.

Table 23 reports the first 10 features selected. We see that these features belong to diverse feature groups (lattice, duration and lexical properties), confirming the importance of involving features that are derived from complementary information sources.

Similarly, the results on the Spanish data are shown in Figure 24. As in the English experiments, the first few features lead to significant R^2 increase and CER decrease, while the remaining features contribute very little. For the most-contributing features shown in Table 6, we observe again that they are composed by diverse features from diverse information resources (lattice, duration and energy). Different from English data, duration and energy-based features show more significant contribution for Spanish experiments, which is also consistent with the individual analysis.

Another observation is that the minimum phone duration contributes to both English and Spanish. This suggests that an abnormal speaking speed tends to indicate an error. Note that phone duration is different from term duration: the former relates to the speaking speed or to strange alignments in ASR errors, while the latter relates more to the length of the search term, and so they are complementary.

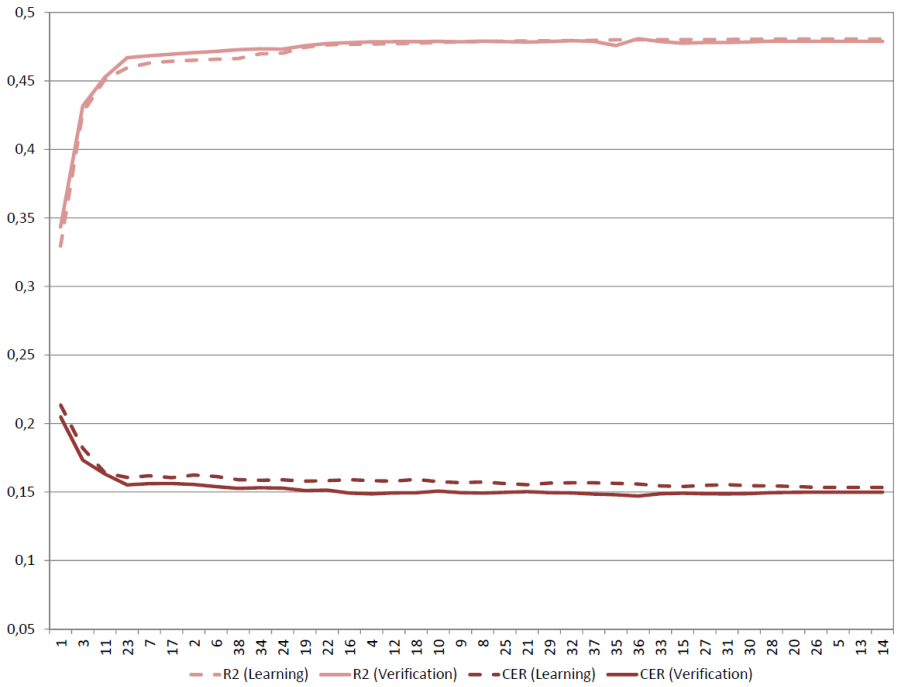

Figure 23: *R^2 and CER results with LR-based incremental feature selection for English data on the learning and verification sets.*

	Learning		Verification	
Feature	R^2	CER%	R^2	CER%
Lattice-based confidence	0.3296	21.34	0.3435	20.48
+Effective FA rate	0.4271	18.20	0.4317	17.32
+Number of phones	0.4511	16.39	0.4528	16.30
+Min phone duration	0.4595	16.06	0.4670	15.51
+Max phone language model score	0.4630	16.19	0.4684	15.61
+Min Levenshtein distance	0.4643	16.04	0.4695	15.62
+Effective occurrence rate	0.4651	16.23	0.4707	15.54
+Mean phone acoustic score	0.4658	16.14	0.4716	15.39
+Min phone energy	0.4664	15.90	0.4728	15.26
+Min energy	0.4698	15.85	0.4735	15.31

Table 5: *The first 10 features selected by LR-based incremental feature analysis for English data. R^2 and CER results on learning and verification sets are also reported.*

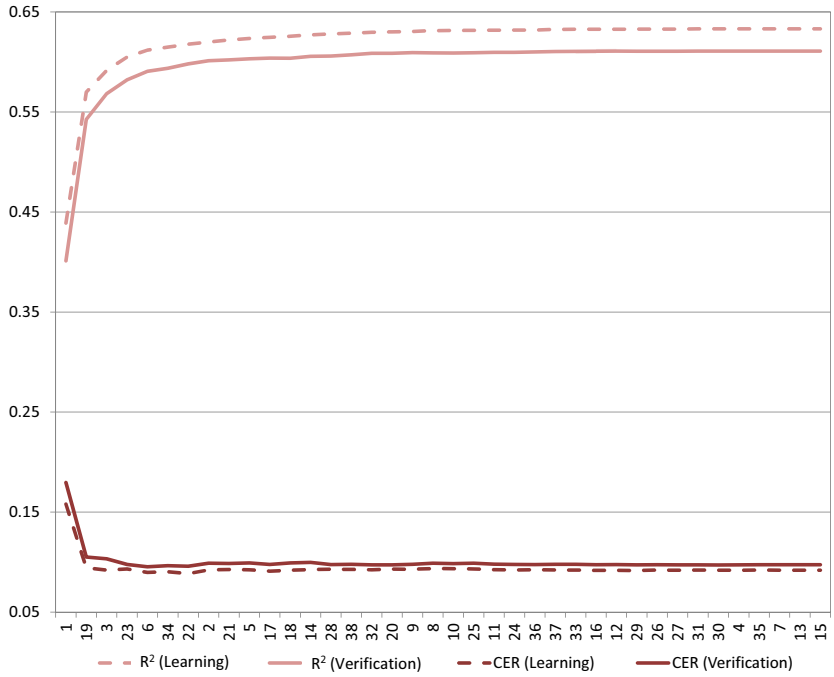

Figure 24: *R^2 and CER results with LR-based incremental feature selection for Spanish data on the learning and verification sets.*

	Learning		Verification	
Feature	R^2	CER%	R^2	CER%
Lattice-based confidence	0.4389	15.80	0.4011	17.96
+Duration	0.5697	9.43	0.5425	10.52
+Effective FA rate	0.5915	9.21	0.5683	10.33
+Min phone duration	0.6049	9.32	0.5821	9.78
+Mean phone acoustic score	0.6118	8.97	0.5906	9.53
+Min energy	0.6148	9.04	0.5936	9.65
+Max phone duration	0.6177	8.85	0.5981	9.59
+Effective occurrence rate	0.6199	9.22	0.6013	9.90
+Vowel speech rate	0.6219	9.26	0.6020	9.86
+Min phone acoustic score	0.6235	9.24	0.6032	9.92

Table 6: *The first* 10 *features selected by LR-based incremental feature analysis for Spanish data. R^2 and CER results on learning and verification sets are also reported.*

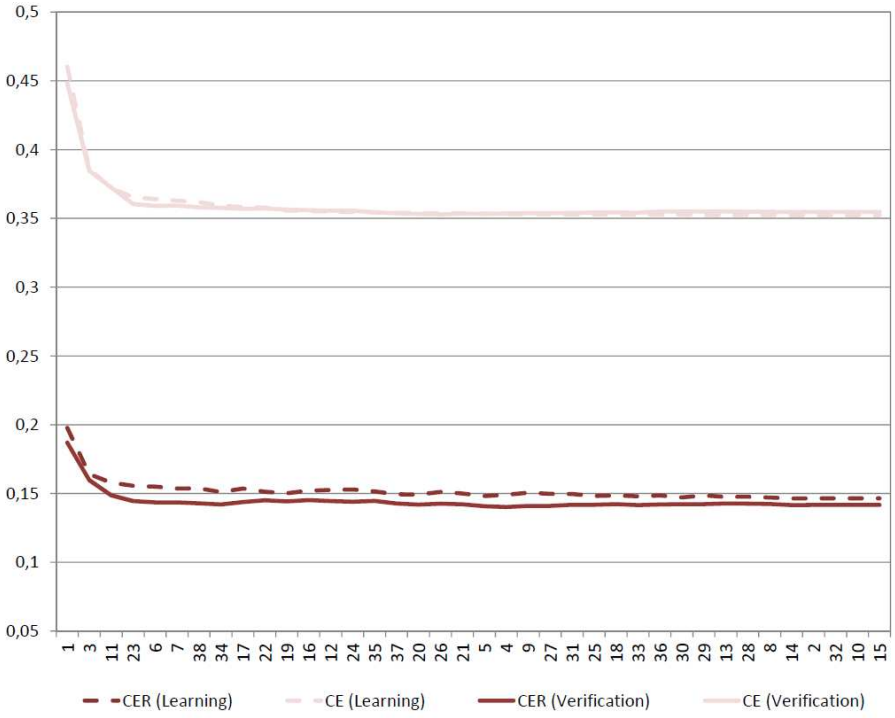

Figure 25: *CE and CER results with LLR-based incremental feature selection for English data on learning and verification sets.*

6.2. Discriminative analysis

In the next experiment we used LLR to perform the incremental selection. The experimental settings are the same as in the regression analysis with LR, except that the selection model and metric are changed to LLR and CE. Again, CER values are presented to show the discriminative power of the selected features.

Figure 25 presents the experimental results on English data, where the four curves represent CE and CER on the learning and verification sets. We observe that the four curves show very similar behavior: with the first 3 features, the CE/CER is substantially reduced and then little further reduction can be obtained by adding more features. The consistency between CE and CER confirms that the cross entropy is directly related to classification performance, and the consistency between the learning and verification sets indicates the learning possesses good generality.

Table 7 reports the first 10 selected features. Comparing with Table 5

	Learning		Verification	
Feature	CE	CER%	CE	CER%
Lattice-based confidence	0.4602	19.76	0.4483	18.70
+Effective FA rate	0.3860	16.42	0.3846	15.97
+Number of phones	0.3718	15.79	0.3724	14.88
+Min phone duration	0.3654	15.56	0.3604	14.45
+Mean phone acoustic score	0.3640	15.50	0.3591	14.36
+Max phone language model score	0.3628	15.37	0.3593	14.36
+Min phone energy	0.3619	15.37	0.3581	14.28
+Min energy	0.3592	15.10	0.3577	14.20
+Min Levenshtein distance	0.3583	15.36	0.3569	14.38
+Max phone duration	0.3578	15.13	0.3572	14.51

Table 7: *The first* 10 *features selected by LLR-based incremental feature analysis for English data. CE and CER results on learning and verification sets are also reported.*

obtained with LR, we see that the first 4 features are exactly the same, although the rest less informative features show differences. This suggests again that keeping a small set of the most useful features is the best approach in obtaining best performance.

The Spanish results are shown in Figure 26. We find the same trend and consistency as in the English experiments: the first few features lead to substantial CE/CER reduction while the remaining features contribute little; the CE and CER curves are highly consistent, and the results on learning and verification sets also exhibit consistency.

Table 8 reports the first 10 selected features. Comparing with the LR results in Table 6, the first 4 features match though the rest of the features show differences.

The results we obtained so far clearly show that the most relevant features obtained by LR and the most discriminative features obtained by LLR largely coincide for STD tasks, so that we can choose any technique to conduct the selection. However, this does not mean LR and LLR are the same: comparing the CER values obtained with LR and LLR, we see that using the LLR method results in lower CER: LLR is a more suitable method for the classification problem than LR.

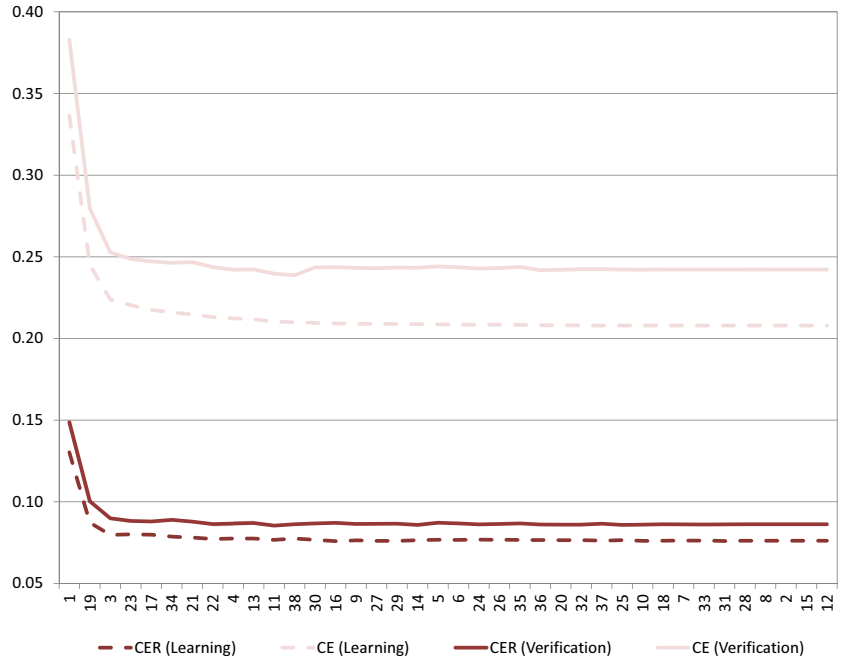

Figure 26: *CE and CER results with LLR-based incremental feature selection for Spanish data on learning and verification sets.*

	Learning		Verification	
Feature	CE	CER%	CE	CER%
Lattice-based confidence	0.3364	13.03	0.3828	14.87
+Duration	0.2451	8.72	0.2795	10.03
+Effective FA rate	0.2238	7.97	0.2527	8.99
+Min phone duration	0.2204	8.01	0.2487	8.83
+Min Levenshtein distance	0.2175	7.98	0.2472	8.79
+Min energy	0.2159	7.87	0.2462	8.89
+Vowel speech rate	0.2148	7.80	0.2467	8.79
+Max phone duration	0.2131	7.72	0.2436	8.64
+Max phone acoustic score	0.2123	7.75	0.2421	8.66
+Number of consonant graphemes	0.2117	7.75	0.2422	8.70

Table 8: *The first* 10 *features selected by LLR-based incremental feature analysis for Spanish data. CE and CER results on learning and verification sets are also reported.*

6.3. Individual and incremental feature analysis comparison

Comparing the individual analysis and incremental analysis, we observe that the best candidates in the incremental selection are not necessarily the most discriminative individuals. For example, the minimum phone duration is selected as the 4th most important feature for English with the LR incremental analysis, but not among the top-10 discriminative features in the individual analysis. This is understandable, as the incremental selection focuses on group discriminative capability, therefore complementarity with other features is more important than individual discrimination. For our example, the minimum phone duration reflects the speaking speed or strange alignments in ASR errors, which is complementary with the features that have been selected: the lattice-based confidence that represents ASR quality, effective FA rate that reflects decision preference, and the number of phones that reflects length of search terms.

7. Feature selection for spoken term detection

In this section we apply the features obtained with the incremental selection to improve spoken term detection. Specifically, the selected features are input into the discriminative confidence estimation framework (see Figure 2), where an MLP model is used to integrate these features and produce the discriminative confidence measure. With feature selection, a less complex MLP can be constructed, which should have better generalization, and of course a lower computational cost.

7.1. MLP training

The MLP models are trained using the same data that are used in the linear regression and logistic linear regression analyses, i.e., a balanced set of detections obtained from the development set. We use a 3-layer MLP (see Figure 3), where the input layer collects all input features, and the output layer contains two units that correspond to hits and FAs respectively. A standard error back-propagation algorithm (Bishop, 1995) is employed to train the model, and K-fold cross-validation ($K = 10$) is employed to select the number of units in the hidden layer.

7.2. Phone recognition experiments

We first report performance of the ASR systems. Since we work with phone-based systems, ASR performance can be evaluated in terms of the

	English	Spanish
PER	40.49%	32.00%
Average lattice density	805	2150

Table 9: *PER and lattice density with the English and Spanish ASR systems.*

phone error rate (PER). Table 9 presents the results. We observe that the Spanish system attains a lower PER than the English system. This can be explained mainly by the difference in speaking style in the databases used (read for Spanish vs. spontaneous for English). This result does not imply that the Spanish STD system is better than the English STD system. In any case, STD performance is impacted by a multitude of factors besides the PER, such as the selection of search terms, the method of confidence estimation, the treatment for overlapped detections and the decision strategy. Among these factors, the quality of lattices is particularly important (see Section 6). We therefore present the lattice density, computed as the average number of nodes per second (Stolcke, 2002) for both languages in Table 9. We find that the Spanish system generates denser lattices than the English system, which imposes a significant impact on quality of the lattice-based confidence, as we will see shortly.

7.3. STD results

We test STD performance based on a discriminative confidence measure derived from features obtained by the incremental selection. Since the LR and LLR-based approaches result in slightly different feature ranks, we present the results with both approaches. Tables 10 and 11 show the ATWV results on the English and Spanish data respectively. We observe that for both languages, the optimal ATWV results are obtained with the $3-5$ most informative features selected by the incremental selection. Including more features does not improve STD. We remark that some of the new features proposed in this paper contribute to the best feature set when evaluating STD performance, e.g., the minimum phone duration and the maximum phone language model score.

Note that the lattice-based confidence itself provides a low ATWV for the Spanish data. This can be attributed to the weak (2-gram) and mismatched (trained on the general domain and tested on the geographical domain) LM. On the one hand, this leads to unreliable lattice-based confidence (which

LR		LLR	
Feature	ATWV	Feature	ATWV
Lattice-based confidence	0.2905	Lattice-based confidence	0.2905
+Effective FA rate	0.2898	+Effective FA rate	0.2898
+Number of phones	0.2911	+Number of phones	0.2911
+Min phone duration	0.2957	+Min phone duration	**0.2957**
+Max phone language model score	**0.2984**	+Mean phone acoustic score	0.2951
+Min Levenshtein distance	0.2947	+Max phone language model score	0.2956
+Effective occurrence rate	0.2947	+Min phone energy	0.2921
+Mean phone acoustic score	0.2957	+Min energy	0.2896

Table 10: *ATWV results on the English data with features incrementally selected based on LR and LLR with the best result in bold font.*

involves LM scores), and on the other hand, this weak LM leads to dense lattices which in turn generates a high number of FAs. Both of the above result in an ATWV lower than that of the English system. The average lattice density has been reported in Table 9.

For the LR-based selection, we observe that choosing the 5 best features provides the best STD performance on the English data, whereas for the Spanish data, choosing the 3 best features is enough to achieve the maximum improvement. Paired t-tests show that, on English data, the best ATWV gain over the single best feature (i.e., lattice-based confidence) is weakly significant ($p < 0.02$), whereas the gain is highly significant ($p < 0.001$) on the Spanish data.

With LLR-based selection, the 4 best features give the best improvement on the English data and the 5 best features do for the Spanish data. Paired t-tests show that the best ATWV gain over the single best feature is highly significant ($p < 0.001$) on the Spanish data, but insignificant for English.

These results demonstrate that both feature selection methods are highly effective in improving STD performance by including on the few most informative features and excluding unuseful features. Although slightly different in terms of ATWV, the LR and LLR analyses resulted in highly consistent feature ranks.

LR		LLR	
Feature	ATWV	Feature	ATWV
Lattice-based confidence	0.0576	Lattice-based confidence	0.0576
+Duration	0.2295	+Duration	0.2295
+Effective FA rate	**0.2343**	+Effective FA rate	0.2343
+Min phone duration	0.2314	+Min phone duration	0.2314
+Mean phone acoustic score	0.2296	+Min Levenshtein distance	**0.2366**
+Min energy	0.2267	+Min energy	0.2320
+Max phone duration	0.2242	+Vowel speech rate	0.2360
+Effective occurrence rate	0.2263	+Max phone duration	0.2236

Table 11: *ATWV results on the Spanish data with features incrementally selected based on LR and LLR with the best result in bold font.*

8. Conclusions

This paper studied various features for STD within the discriminative confidence estimation framework. Two analysis tools based on linear regression and logistic linear regression are employed to study the contribution of these features to STD individually and as a group. The experiments were conducted on two databases: one contains English meeting speech and the other contains Spanish read speech. Our analysis shows that for both the English data and the Spanish data, the lattice-based confidence, the effective FA rate and the minimum phone duration are generally the most important, and the best feature set is composed of features derived from diverse sources (ASR decoding, duration and lexical properties). Although based on different criteria, the LR and LLR analyses lead to highly consistent feature ranks. This indicates that for STD, the most relevant features are also the most discriminative. In spite of the complexity in methodology and data, the candidate features proposed have been demonstrated to deliver significant performance improvements when compared to a baseline using the single best feature.

Future work involves more advanced analyzing approaches, especially various approaches to automatic relevance detection (ARD), and sparse discriminative analysis (SDA). Extending the incremental search to more comprehensive search approaches such as evolutionary algorithms might be another interesting direction.

9. Acknowledgements

This work has been partially supported by project PriorSPEECH (TEC2009-14719-C02-01) from the Spanish Ministry of Science and Innovation and by project MAV2VICMR (S2009/TIC-1542) from the Community of Madrid.

References

Akbacak, M., Vergyri, D., Stolcke, A., March 2008. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. In: Proc. ICASSP'08. Las Vegas, Nevada, USA, pp. 5240–5243.

Almuallim, H., Dietterich, T. G., 1991. Learning with many irrelevant features. In: Proceedings of the National Conference on Artificial Intelligence. pp. 547–552.

Almuallim, H., Dietterich, T. G., 1994. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence 69, 279–305.

Ayed, Y. B., Fohr, D., Haton, J. P., Chollet, G., November 2002. Keyword spotting using support vector machines. In: Proc. International Conference on Text, Speech and Dialogue (TSD). Brno, Czech Republic, pp. 285–292.

Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y., Guyon, I., Elisseeff, A., March 2003. Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research 3, 1183–1208.

Ben-Bassat, M., 1982. Pattern recognition and reduction of dimensionality. Handbook of Statistics II 1.

Bergen, Z., Ward, W., September 1997. A senone based confidence measure for speech recognition. In: Proc. Eurospeech 97. Rhodes, Greece, pp. 819–822.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., Song, M., March 2003. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research 3, 1229–1243.

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

Boersma, P., Weenink, D., 2007. Praat: doing phonetics by computer. University of Amsterdam, Spuistraat 210, Amsterdam, Holland.
URL http://www.fon.hum.uva.nl/praat/

Breiman, L., Friedman, J. H., Olshen, R. A., Stone., C. J., 1984. Classification and Regression Trees. Wadsworth and Brooks.

Can, D., Cooper, E., Sethy, A., White, C., Ramabhadran, B., Saraçlar, M., April 2009. Effect of pronunciations on OOV queries in spoken term detection. In: Proc. ICASSP'09. Taipei, Taiwan, pp. 3957–3960.

Chan, C., Lee, L., September 2010. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. In: Proc. Interspeech'10. Makuhari, Chiba, Japan, pp. 693–696.

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

Chen, C., Lee, H., Yeh, C., Lee, L., September 2010. Improved spoken term detection by feature space pseudo-relevance feedback. In: Proc. Interspeech'10. Makuhari, Chiba, Japan, pp. 1672–1675.

Cox, S., Rose, R., May 1996. Confidence measures for the SWITCHBOARD database. In: Proc. ICASSP'96. Vol. 1. Atlanta, Georgia, USA, pp. 511–514.

Deligne, S., Yvon, F., Bimbot, F., September 1995. Variable-length sequence matching for phonetic transcription using joint multigrams. In: Proc. Eurospeech'95. Madrid, Spain, pp. 2243–2246.

Duda, R. O., Hart, P. E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern classification. Wiley, New York.

Forman, G., March 2003. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., Haussler, D., May 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16 (10), 906–914.

Gadde, V. R. R., 2000. Modeling word duration for better speech recognition. In: Proceedings of the Speech Transcription Workshop. pp. 1–4.

Gillick, L., Ito, Y., Young, J., April 1997. A probabilistic approach to confidence estimation and evaluation. In: Proc. ICASSP'97. Munich, Bavaria, Germany, pp. 879–882.

Goldwater, S., Jurafsky, D., Maning, C. D., 2009. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication 52 (3), 181–200.

Good, I. J., 1965. The Estimation Of Probabilities: An Essay on Modern Bayesian Methods. MIT, Cambridge.

Guyon, I., Elisseeff, A., March 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.
URL `http://portal.acm.org/citation.cfm?id=944919.944968`

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., March 2002. Gene selection for cancer classification using support vector machines. Journal of Machine Learning 46, 389–422.

Hain, T., Burget, L., Dines, J., Garau, G., Karafiát, M., Lincoln, M., Vepa, J., Wan, V., 2006. The AMI meeting transcription system: Progress and performance. In: Machine Learning for Multimodal Interaction. Vol. 4299/2006. Springer Berlin/Heidelberg, pp. 419–431.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The weka data mining software: An update. SIGKDD Explorations 11 (1), 10–18.

Hall, M. A., April 1999. Correlation-based feature selection for machine learning. Ph.D. thesis, Department of Computer Science, Waikato University, New Zealand.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hauptmann, A. G., Jones, R. E., Seymore, K., Slattery, S. T., Witbrock, M. J., Siegler, M. A., February 1998. Experiments in information retrieval from spoken documents. In: Proc. DARPA Workshop on Broadcast News Transcription and Understanding. Lansdowne VA, pp. 175–181.

Hellevik, O., 2009. Linear versus logistic regression when the dependent variable is a dichotomy. Quality & Quantity 43, 59–74.

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

Jansen, A., Church, K., Heřmanský, H., September 2010. Towards spoken term discovery at scale with zero resources. In: Proc. Interspeech'10. Makuhari, Chiba, Japan, pp. 1676–1679.

Jiang, H., 2005. Confidence measures for speech recognition: A survey. Speech Communication 45 (4), 455–470.

Kamppari, S. O., Hazen, T. J., June 2000. Word and phone level acoustic confidence scoring. In: Proc. ICASSP'00. Vol. 3. Istanbul, Turkey, pp. 1799–1802.

Kao, J. T., Zweig, G., Nguyen, P., May 2011. Discriminative duration modeling for speech recognition with segmental conditional random fields. In: Proc. ICASSP'11. Prague, Czech Republic, pp. 4476–4479.

Kemp, T., Schaaf, T., September 1997. Estimating confidence using word lattices. In: Proc. Eurospeech'97. Rhodes, Greece, pp. 827–830.

Kira, K., Rendell, L. A., 1992a. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the National Conference on Artificial intelligence. pp. 129–134.

Kira, K., Rendell, L. A., 1992b. A practical approach to feature selection. In: Proceedings of the International Workshop on Machine learning. pp. 249–256.

Kohavi, R., John, G. H., December 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

Koller, D., Sahami, M., July 1996. Toward optimal feature selection. In: Proc. International Conference on Machine Learning. Bari, Italy, pp. 284–292.

Kononenko, I., 1994. Estimating attributes: Analysis and extensions of relief. In: Proceedings of European Conference on Machine Learning. pp. 171–182.

Langley, P., Iba, W., Thompson, K., 1992. An analysis of bayesian classifiers. In: Proceedings of the National Conference on Artificial Intelligence. pp. 223–228.

Liaw, A., Wiener, M., December 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Logan, B., Moreno, P., Thong, J.-M. V., Whittaker, E., October 2000. An experimental study of an audio indexing system for the web. In: Proc. ICSLP'00. Vol. 2. Beijing, China, pp. 676–679.

Mamou, J., Ramabhadran, B., September 2008. Phonetic query expansion for spoken document retrieval. In: Proc. Interspeech'08. Brisbane, Australia, pp. 2106–2109.

Mamou, J., Ramabhadran, B., Siohan, O., July 2007. Vocabulary independent spoken term detection. In: Proc. ACM-SIGIR'07. Amsterdam, The Netherlands, pp. 615–622.

Manos, A., Zue, V., April 1997. A segment-based wordspotter using phonetic filler models. In: Proc. ICASSP'97. Vol. 2. Munich, Bavaria, Germany, pp. 899–902.

Mathan, L., Miclet, L., May 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, Ontario, Canada, pp. 93–96.

Meng, S., Yu, P., Seide, F., Liu, J., December 2007. A study of lattice-based spoken term detection for Chinese spontaneous speech. In: Proc. ASRU'07. Kyoto, Japan, pp. 635–640.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., Nadeu, C., September 1993. Albayzin speech database: Design of the phonetic corpus. In: Proc. Eurospeech'93. Berlin, Germany, pp. 653–656.

Motlicek, P., Valente, F., Garner, P., September 2010. English spoken term detection in multilingual recordings. In: Proc. Interspeech'10. Makuhari, Chiba, Japan, pp. 206–209.

Neti, C. V., Roukos, S., Eide, E., April 1997. Word-based confidence measures as a guide for stack search in speech recognition. In: Proc. ICASSP'97. Munich, Bavaria, Germany, pp. 883–886.

NIST, September 2006. The spoken term detection (STD) 2006 evaluation plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10th Edition.
URL `http://www.nist.gov/speech/tests/std`

NIST, July 2013. The OpenKWS13 Evaluation Plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 1st Edition.
URL `http://www.nist.gov/itl/iad/mig/openkws13.cfm`

Ou, J., Chen, K., Wang, X., Li, Z., November 2001. Utterance verification of short keywords using hybrid neural-network/HMM approach. In: Proc. International Conference on Info-tech and Info-net (ICII). Beijing, China, pp. 671–676.

Parada, C., Sethy, A., Ramabhadran, B., March 2010. Balancing false alarms and hits in spoken term detection. In: Proc. ICASSP'10. Dallas, Texas, USA, pp. 5286–5289.

Parlak, S., Saraçlar, M., March 2008. Spoken term detection for Turkish broadcast news. In: Proc. ICASSP'08. Las Vegas, Nevada, USA, pp. 5244–5247.

Pinto, J., Szöke, I., Prasanna, S., Heřmanský, H., July 2008. Fast approximate spoken term detection from sequence of phonemes. In: Proc. ACM-SIGIR'08. Singapore, pp. 28–33.

Rohlicek, J. R., Russell, W., Roukos, S., Gish, H., May 1989. Continuous hidden Markov modeling for speaker-independent word spotting. In: Proc. ICASSP'89. Glasgow, UK, pp. 627–630.

Saeys, Y., Inza, I., Larrañaga, P., October 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.

Saraçlar, M., Sproat, R., May 2004. Lattice-based search for spoken utterance retrieval. In: Proc. HLT-NAACL 2004. Boston, USA, pp. 129–136.

Schaaf, T., Kemp, T., April 1997. Confidence measures for spontaneous speech recognition. In: Proc. ICASSP'97. Munich, Bavaria, Germany, pp. 875–878.

Shafran, Z., Roark, B., Fisher, S., December 2006. OGI spoken term detection system. In: Proc. NIST spoken term detection workshop (STD 2006). Gaithersburg, Maryland, USA, pp. 1–15.

Siu, M., Gish, H., Richardson, F., September 1997. Improved estimation, evaluation and applications of confidence measures for speech recognition. In: Proc. Eurospeech'97. Rhodes, Greece, pp. 831–834.

Stolcke, A., 2002. SRILM - an extensible language modeling toolkit. In: Proc. of Interspeech'02. Denver, USA, pp. 901–904.

Sudoh, K., Tsukada, H., Isozaki, H., September 2006. Discriminative named entity recognition of speech data using speech recognition confidence. In: Proc. ICSLP'06. Pittsburgh, USA, pp. 1153–1156.

Sukkar, R. A., Wilpon, J. G., April 1993. A two pass classifier for utterance rejection in keyword spotting. In: Proc. ICASSP'93. Vol. 2. Minneapolis, MN, USA, pp. 451–454.

Szöke, I., Burget, L., Černocký, J., Fapšo, M., December 2008a. Sub-word modeling of out of vocabulary words in spoken term detection. In: Proc. SLT'08. Goa, India, pp. 273–276.

Szöke, I., Fapšo, M., Karafiát, M., Burget, L., Grézl, F., Schwarz, P., Glembek, O., Matějka, P., Kontár, S., Černocký, J., December 2006. BUT system for NIST STD 2006 - English. In: Proc. NIST spoken term detection workshop (STD 2006). Gaithersburg, Maryland, USA, pp. 1–26.

Szöke, I., Fapšo, M., Karafiát, M., Burget, L., Grézl, F., Schwarz, P., Glembek, O., Matějka, P., Kopecký, J., Černocký, J., 2008b. Spoken term detection system based on combination of LVCSR and phonetic search. In:

Machine Learning for Multimodal Interaction. Vol. 4892/2008 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 237–247.

Tejedor, J., April 2009. Contributions to keyword spotting and spoken term detection for information retrieval in audio mining. Ph.D. thesis, Human Computer Technology Laboratory, Universidad Autónoma de Madrid.

Tejedor, J., Toledano, D. T., Bautista, M., King, S., Wang, D., Colás, J., September 2010. Augmented set of features for confidence estimation in spoken term detection. In: Proc. Interspeech'10. Makuhari, Chiba, Japan, pp. 701–704.

Tejedor, J., Wang, D., Frankel, J., King, S., Colás, J., November 2008. A comparison of grapheme and phoneme-based units for Spanish spoken term detection. Speech Communication 50 (11-12), 980–991.

Thambiratmann, K., Sridharan, S., January 2007. Rapid yet accurate speech indexing using dynamic match lattice spotting. IEEE Transactions on Audio, Speech, and Language Processing 15 (1), 346–357.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society 73 (3), 273–282.

Torkkola, K., March 2003. Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research 3, 1415–1438.

Tusher, V. G., Tibshirani, R., Chu, G., April 2001. Significance analysis of microarrays applied to the ionizing radiation response. National Academy of Sciences 98 (5), 5116–5121.

Vergyri, D., Shafran, I., Stolcke, A., Gadde, R. R., Akbacak, M., Roark, B., Wang, W., August 2007. The SRI/OGI 2006 spoken term detection system. In: Proc. Interspeech'07. Antwerp, Belgium, pp. 2393–2396.

Vergyri, D., Stolcke, A., Gadde, R. R., Wang, W., December 2006. The SRI 2006 spoken term detection system. In: Proc. NIST spoken term detection workshop (STD 2006). Gaithersburg, Maryland, USA, pp. 1–15.

Wallace, R., Vogt, R., Baker, B., Sridharan, S., March 2010. Optimising figure of merit for phonetic spoken term detection. In: Proc. ICASSP'10. Dallas, Texas, USA, pp. 5298–5301.

Wallace, R., Vogt, R., Sridharan, S., August 2007. A phonetic search approach to the 2006 NIST spoken term detection evaluation. In: Proc. Interspeech'07. Antwerp, Belgium, pp. 2385–2388.

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

Wang, D., Evans, N. W. D., Troncy, R., King, S., May 2011. Handling overlaps in spoken term detection. In: Proc. ICASSP'11. Prague, Czech Republic, pp. 5656–5659.

Wang, D., Frankel, J., Tejedor, J., King, S., March 2008. A comparison of phone and grapheme-based spoken term detection. In: Proc. ICASSP'08. Las Vegas, Nevada, USA, pp. 4969–4972.

Wang, D., King, S., Frankel, J., September 2009a. Stochastic pronunciation modelling for spoken term detection. In: Proc. Interspeech'09. Brighton, UK, pp. 2135–2138.

Wang, D., King, S., Frankel, J., Bell, P., September 2009b. Term-dependent confidence for out-of-vocabulary term detection. In: Proc. Interspeech'09. Brighton, UK, pp. 2139–2142.

Wang, D., Tejedor, J., King, S., Frankel, J., March 2012. Term-dependent confidence normalisation for out-of-vocabulary spoken term detection. Journal of Computer Science and Technology 27 (2), 358–375.

Weintraub, M., Beaufays, F., Rivlin, Z., Konig, Y., Stolcke, A., April 1997. Neural-network based measures of confidence for word recognition. In: Proc. ICASSP'97. Munich, Bavaria, Germany, pp. 887–890.

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M., March 2003. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research 3, 1439–1461.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., March 2006. The HTK v3.4 Book. Engineering Department, Cambridge University.

Yu, L., Liu, H., December 2004. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224.

Zhang, R., Rudnicky, A. I., September 2001. Word level confidence annotation using combinations of features. In: Proc. Eurospeech'01. Aalborg, Denmark, pp. 2105–2108.

10. Vitae

Javier Tejedor received a B.Sc. in Computer Engineering, M.Sc. in Computer and Telecommunication Engineering and Ph.D. degree in Computer and Telecommunication Engineering in 2002, 2005 and 2009 respectively from the Universidad Autónoma de Madrid. He is currently with the Human Computer Technology Laboratory (HCTLab) in the School of Computer Engineering and Telecommunication in that same university and is an assistant professor in the School of Computer Engineering and Telecommunication in the same university. His main interests are speech indexing and retrieval, spoken term detection and large vocabulary continuous speech recognition.

Doroteo T. Toledano received the Telecommunication Engineering degree from the Universidad Politécnica de Madrid, Spain, in 1997, obtaining the best academic records of his class. In 2001, he completed the Ph.D. degree in telecommunication engineering from the same university, receiving a Ph.D. Dissertation Award from the Spanish Association of Telecommunication Engineers. He was with the Speech Technology Division of Telefónica R&D from 1994 to 2001. From 2001 to 2002, he was with the Spoken Language Systems Group at the Massachusetts Institute of Technology, Laboratory for Computer Science as a Postdoctoral Research Associate. After another short period at the Speech Technology Division of Telefónica R&D he moved in 2004 to the Universidad Autónoma de Madrid (Spain) where he is currently Associate Professor. His current research interests include acoustic modeling, speaker and language recognition, and multimodal biometrics. He has over 70 scientific publications in these fields including journals and conference proceedings. He has also served as a member of the scientific committee of several international conferences as well as a reviewer for several journals such as IEEE Transactions on Audio, Speech and Language Processing, IEEE Signal Processing Letters, Speech Communication and Computer Speech and Language.

Dong Wang received the B.Sc. and M.Sc. in computer science at Tsinghua Univ. in 1999 and 2002, and then worked for Oracle China in 2002-2004 and IBM China in 2004-2006. He joined CSTR, University of Edinburgh in 2006 as a research fellow and PhD student supported by a Marie Curie fellowship, from where he received his Ph.D. in 2010. From 2010 to 2011 he worked in EURECOM where he extended his research to speech processing, spoken document retrieval and machine learning in general. He is now working in Nuance USA as a senior research engineer.

Simon King (M'95 - SM'08) received the M.A.(Cantab) and M.Phil. degrees in engineering from the University of Cambridge, Cambridge, U.K., in 1992 and 1993 and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 1998. He is Professor of Speech Processing at the University of Edinburgh and his interests include speech synthesis, recognition, and signal processing. Prof. King has served on the ISCA SynSIG committee, co-organizes the Blizzard Challenge, was recently an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing, is on the IEEE SLTC and the editorial board of Computer Speech and Language.

José Colás is professor within the Computer Architecture and Technology area since 2002. He received his Bachelor degree in Telecommunication Engineering from the Universidad Politécnica de Madrid in 1990 and the Ph.D. degree in Telecommunications from the same university in 1999. In 1993 his group received the *Reina Sofia* award for a research trajectory focused on technologies for the disabled. In 2001 he founded the Human Computer Technology Laboratory (HCTLab) at the Universidad Autónoma de Madrid. This group received in 2003 the *Infanta Cristina* award for their research related to the new technologies for disability focused on mobile devices. He is the head of the "Multimodal Interaction oriented to Disabled people" research at the HCTLab.