CONTENTS

LIST OF FIGURES

Large Vocabulary Continuous Speech Recognition: Advances and Applications

Jean-Luc Gauvain and Lori Lamel

Abstract—

The last decade has witnessed substantial advances in speech recognition technology, which when combined with the increase in computational power and storage capacity, has resulted in a variety of commercial products already or soon to be on the market. In this paper we review the state-of-the-art in core technology in large vocabulary continuous speech recognition, with a view towards highlighting recent advances. We then highlight issues in moving towards applications, discussing system efficiency, portability across languages and tasks, enhancing the system output by adding tags and non-linguistic information. Current performance in speech recognition and outstanding challenges for three classes of applications: dictation, audio indexation and spoken language dialog systems are discussed.

Keywords— **Speech recognition, spoken language systems, dictation, large vocabulary, speaker-independent continuous speech recognition, acoustic modeling, model adaptation, portability, multilinguality**

I. Introduction

This paper overviews recent advances in state-of-the-art laboratory speech recognition systems, and explores application domains made possible by technological progress. Only a few years ago speech recognition was primarily associated with a limited number of applications: small vocabulary isolated word recognition (IWR) or phrases, mid-sized vocabulary domain specific spoken language systems, and dictation systems (often for specific user groups). For the last decade large vocabulary, continuous speech recognition (LVCSR) has been one of the focal areas of research in speech recognition, serving as a test bed to evaluate models and algorithms.

The core technology developed for LVCSR can be used for applications other than general dictation systems, it also serves as the basis for less demanding applications such as voice-interactive database access or limited-domain dictation, as well as more demanding tasks such as the transcription of broadcast data. Progress in speech recognition can also boost other spoken language technologies such as speaker and language identification which rely on the same modeling techniques.

With the exception of the inherent variability of telephone channels, in most applications it can be assumed that the speech is produced in relatively stable environmental (background acoustic conditions) and in the case of dictation, is spoken with the purpose of being transcribed by the machine. A major advance is the ability of todays laboratory systems to deal with non-homogeneous data as is exemplified by broadcast data: changing speakers, languages, backgrounds, topics. This capability has been enabled by advances in techniques for robust signal processing and normalization; improved training techniques which can take advantage of very large audio and textual corpora; algorithms for audio segmentation; unsupervised acoustic model adaptation; efficient decoding with long span language models; ability to use much larger vocabularies than in the past - 64 k words or more is common to reduce errors due to out-of-vocabulary words; and by the adoption of assessment-driven technology development methodology largely fostered by the US DARPA efforts.

In this paper we restrict our attention essentially to large vocabulary continuous speech recognition. However developing systems based on this technology goes far beyond automatic speech recognition, and involves other domains such as human factors and user interface design, natural language understanding, generation and synthesis as well integration with the back-end (database) or user (often already existing) infrastructure. There are many issues such as efficiency and costs considerations of final product (central server vs. distributed) that are not discussed. Moving towards real-world applications means building usable systems which involves reconsidering many design issues such as signal capture, noise and channel compensation, and rejection capability, while taking into account limitations in computational resources. The difficulties and costs of adapting existing technology to new languages or new applications must also be evaluated.

In the next section we review the state-of-the-art in large vocabulary continuous speech recognition, focusing on what is in the public domain which often implies laboratory systems. The highlighted techniques were chosen based on experimental results obtained in different laboratories on publicly available data using state-of-the-art systems. While we attempt to generalize the description, some details pertain to LIMSI systems. Section IV discusses three main classes of applications: dictation, audio indexing and dialog systems; as well as some of what we consider to be outstanding challenges for speech recognition in the context of these applications.

II. Core technology for LVCSR

Speech recognition is primarily concerned with transcribing a speech signal as a sequence of words. Most of today's best performing systems are based on a statistical model of speech generation. From this point of view, speech is assumed to be generated by a language model which provides estimates of $\Pr(w)$ for all word strings w independently of the observed signal, and a model of the acoustic channel encoding the message w in the signal x, which is represented by a probability density function $f(x|w)$. The speech decoding problem is to maximize the *a posteriori* probability of w, which is equivalent to maximizing the product $\Pr(w)f(x|w)$. The basic principles on which most state-of-the-art continuous speech recognizers are based have been known for many years, and include the application of information theory to speech recognition [8], [71], the use of a spectral representation of the speech signal [33], [34], the use of dynamic programming for decoding [153], [154], and the use of

Jean-Luc Gauvain and Lori Lamel are with the LIMSI-CNRS, France. E-mail: gauvain@limsi.fr and lamel@limsi.fr

context-dependent acoustic models [24], [90], [140]. In spite of this considerable progress has been made in recent years, particularly in acoustic modeling and decoding. Much of this progress can be linked to the availability of large speech and text corpora and simultaneous advances made in computational means and storage, which have facilitated the implementation of more complex models and algorithms.

A. Acoustic-Phonetic Modeling

Most state-of-the-art LVCSR systems make use of hidden Markov models (HMM) for acoustic modeling [166]. Other approaches include segment based models [59], [117], [172] and neural networks [2], [68] to estimate acoustic observation likelihoods. However except for the acoustic likelihood estimation, all systems make use of the HMM framework to combine linguistic and acoustic information in a single network representing all possible sentences.

For HMM based systems, acoustic modeling consists of modeling the probability density function of a sequence of acoustic feature vectors. The acoustic features are chosen so as to reduce model complexity while trying to keep the relevant information (i.e., the linguistic information for the speech recognition problem). Most recognition systems use short-term cepstral features based either on a Fourier transform or a linear prediction model. The two most popular sets of features are cepstrum coefficients obtained with an MFCC [28] analysis or with a PLP [67] analysis. In both cases a Mel scale short term power spectrum is estimated on a fixed window (usually in the range of 20 to 30 ms), with the most commonly used frame rate being 10ms. To get the MFCC cepstrum coefficients a cosine transform is applied to the log power spectrum, whereas a root-LPCC analysis is used to obtain the PLP cepstrum coefficients. Both set of features have been successfully used, but PLP analysis has been found for some systems to be more robust in presence of background noise [78], [163]. Finding the optimal tuning, which may be dependent on the language or the channel conditions, can result in slight performance improvements.

As an example, the LIMSI front end used to transcribe broadcast news data produces a feature vector containing 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8 kHz band (or 0-3.5 kHz for telephone data) every 10 ms. For each 30 ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. These cepstral coefficients are normalized using cepstral mean removal [41] and variance normalization. Each resulting cepstral coefficient therefore has a zero mean and unity variance.

Most recognition systems use acoustic units corresponding to phonemic or phonetic units (or phones in context). However it is certainly possible to perform speech recognition without use of a phonemic lexicon, either by use of “word models” or a different mapping such as the fenonic lexicon [10]. Compared to word models, subword units reduce the number of parameters, enable cross word modeling and facilitate porting to new vocabularies. Fenones offer the additional advantage of automatic training, but lack the ability to include *a priori* linguistic knowledge. Context-dependent (CD) phone models are today the most commonly used acoustic units for LVCSR. Compared to larger units such as *diphones, demisyllables* or *syllables*, a large spectrum of contextual dependencies can be implemented for CD models associated with backoff mechanisms to model infrequent contexts. Various types of contexts have been investigated from a single phone context (right- or left-context), left and right-context (triphone), generalized triphones [90], position-dependent triphones (cross-word and within word triphones), function word triphones, and quinphones [162]. While different approaches are used to select the phone contexts (often based on frequency of occurrence or phonetic decision trees), the optimal set of modeled contexts is usually the result of a tradeoff between resolution and robustness, and is highly dependent on the available training data. This optimization is generally done by minimizing the recognizer error rate on development data. In fact, more than the number of CD phone models, what is really important is to match the total number of model parameters to the amount of available training data. A powerful technique to keep the models trainable without sacrificing model resolution is to take advantage of the state similarity among different models of a given phone by tying the HMM state distributions. This basic idea is used in most current systems although there are slight differences in the implementation and in the naming of the resulting clustered states (*senones* [69], *genones* [30], *PELs* [13], *tied-states* [170]). Numerous ways of tying HMM parameters have been investigated [150], [165] in order to overcome the sparse training data problem and to reduce the need for distribution smoothing techniques.

In practice both agglomerative clustering and divisive clustering have been found to yield model sets with comparable performance. Divisive decision tree clustering is particularly interesting when there are a very large number of states to cluster since it is at the same time both faster and is more robust than a bottom-up greedy algorithm, and therefore much easier to tune. In addition, HMM state tying based on decision tree clustering has the advantage of providing a means to build models for unseen contexts, i.e., those contexts which do not occur in the training data [70], [169]. The set of questions typically concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones [111].

Many state-of-the-art recognizers make use of continuous density HMM with Gaussian mixture for acoustic modeling. The main advantage continuous density modeling offers over discrete or semi-continuous (or tied-mixture) observation density modeling is that the number of parameters used to model an HMM observation distribution can easily be adapted to the amount of available training data associated to this state. As a consequence, high precision modeling can be achieved for highly frequented states without the explicit need of smoothing techniques for the densities of less frequented states. Discrete and semi-continuous modeling use a fixed number of parameters to represent a given observation density and therefore cannot achieve high precision without the use of smoothing techniques or tying techniques mentioned above.

The choice of the model structure is highly dependent on the constraints of the application such as limitations on available memory or computational capacity.

It is fairly common practice to use separate male and female models to more accurately model the speech data. The sex-dependent models are often obtained from speaker-independent seed models using Maximum *A Posteriori* estimators [55], or may be trained on the independent data subsets if sufficient training data are available.

B. Lexical Representation

Lexical modeling provides the link between the lexical entries (usually words) used by the language model and the acoustic models, with each lexical entry being described as a sequence of elementary units. Experience has shown that systematic lexical design can improve system performance [82]. Lexical design entails two main parts - selection of the vocabulary items and representation of the pronunciation entry using the basic units of the recognition system. A common way of selecting a recognition vocabulary is to measure the out-of-vocabulary (OOV) rate on development data. Judicious selection of the development data is important in order to ensure high lexical coverage on the test material. The best lexical coverage may be obtained by selecting the vocabulary using only a subset of the training data (such as the most recent data or data on a given topic) instead of using all the available data [20], [53]. On average, each OOV word causes more than a single error, with rates of 1.6 to 2.0 additional errors reported [119]. An obvious way to reduce the error rate due to OOVs is to increase the size of the lexicon. Increasing the lexicon size to 64 k or more words has been shown to improve performance, despite the potential of increased confusability of the lexical entries [53], so in contradiction to the widely held belief, larger vocabulary does not imply higher word error rates if a proper language model is used.

For LVCSR, the lexical unit of choice is usually phonemes or phoneme-like units, specific for the language. For example, the LIMSI phone set for American English has 46 units, with 45 for British English, 35 for French, 49 for German, 26 for Spanish, and 36 for Mandarin (to which tones may be added). In generating pronunciation baseforms, most lexicons include standard pronunciations and do not explicitly represent allophones. This representation is chosen as most allophonic variants can be predicted by rules, and their use is optional. More importantly, there often is a continuum between different allophones of a given phoneme and the decision as to which occurred in any given utterance is subjective. By using a phonemic representation, no hard decision is imposed, and it is left to the acoustic models to represent the observed variants in the training data. Several efforts to automatically learn and generate word pronunciations have been investigated [21], [26], [40], [129], [149]. To the best of our knowledge such approaches, while promising, have to date, given only small performance improvements even when trained with manual transcriptions [130].

There are a variety of words for which frequent alternative pronunciation variants are observed, and these variants are not due to allophonic differences. One common example is the suffix *-ization* which can be pronounced with a diphthong or a schwa. Alternate pronunciations are also needed for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as *excuse, record, produce*.

Fast speakers tend to poorly articulate unstressed syllables (and sometimes skip them completely), particularly in long words with sequences of unstressed syllables. Although such long words are typically well recognized, often a nearby function word is deleted. To reduce these kinds of errors, alternate pronunciations for long words can authorize schwa deletion or syllabic consonants in unstressed syllables. Phonological rules have been proposed to account for some of the phonological variations observed in fluent speech [116]. The principle behind the phonological rules is to modify the phone network to take into account such variations [26], [56], [85]. These rules can be optionally applied during training and recognition. Using phonological rules during training results in better acoustic models, as they are less “polluted” by wrong transcriptions. Their use during recognition reduces the number of mismatches. The same mechanism can also be used to handle liaisons, mute-e, and final consonant cluster reduction for French [52].

C. Language Modeling

Language models are used to model regularities in natural language [135]. The most popular methods are statistical *n*-gram models which attempt to capture the syntactic and semantic constraints by estimating the probability of a word in a sentence given the preceding *n*-1 words. Different approaches have been investigated to smooth the estimates of the probabilities of rare *n*-grams [22], [79]. The most common is approach is to apply a backoff mechanism [76] relying on a lower order *n*-gram when there is insufficient training data, providing a means of modeling unobserved *n*-grams. Another advantage of the backoff mechanism is that LM size can be arbitrarily reduced by relying more on the backoff component, obtained by simply increasing the minimum number of required *n*-gram observations needed to include the *n*-gram in the model. This property can also be used to reduce computational requirements. While bigram and trigram LMs are most widely used, small improvements have been reported with the use of longer span *4*-grams [9], [162] and *5*-grams [97] or class *5*-grams [137]. Language models are typically compared by measuring the likelihood of a set of development texts.

Given a large corpus of texts (or transcriptions) it may seem relatively straightforward to construct *n*-gram language models. Most of the steps are pretty standard and make use of tools that count word and word sequence occurrences [25]. The main differences arise in the choice of the vocabulary and in the definition of words, such as the treatment of compound words or acronyms, and the choice of the backoff strategy. There is, however, a significant amount of effort needed to process the texts before they can be used.

One motivation for normalization is to reduce lexical variability so as to increase the coverage for a fixed size task vocabulary. Normalization decisions are generally language-specific. For example, some standard processing steps include the expansion of numerical expressions, treatment of isolated letters and letter sequences, and optionally elimination of case distinction. Further semi-automatic processing is necessary to correct frequent errors inherent in the texts, and the expansion of abbreviations and acronyms. The error correction consists primarily of correcting obvious misspellings. Better language models can be obtained by using texts transformed to be closer to the observed reading style, where the transformation rules and corresponding probabilities are automatically derived by aligning prompt texts with the transcriptions of the acoustic data. For example, the number 150 may be pronounced as “one hundred fifty” or “one hundred and fifty”. Similarly, 1/8 may be spoken as “one eighth” or “an eighth” [53].

There sometimes is a conflicting need for sufficient amounts of text data to estimate LM parameters and assuring that the data is representative of the task. It is also common that different types of LM training material are available in differing quantities, that need to be combined. Combining sources requires that common normalizations are carried out. One easy way to combine training material from different sources is to train a language model per source and to interpolate them. The interpolation weights can be directly estimated on some development data with the EM algorithm. An alternative is to simply merge the n-gram counts and train a single language model on these counts. If some data sources are more representative than others for the task, the n-gram counts can be empirically weighted to minimize the perplexity on a set of development data. While this can be effective, it has to be done by trial and error and cannot easily be optimized. In addition, weighting the n-gram counts can pose problems in properly estimating the backoff coefficients.

Word class-based language models can be used to reduce the dependency on the training data, particularly when there is no *a priori* reason to believe that any member of the class is more likely than another. This technique is often used in spoken language dialog systems for common items such as locations, dates and times.

D. Decoding

The main challenge for LVCSR decoding problem is the design of an efficient search algorithm to deal with the huge search space obtained by combining the acoustic and language models. Strictly speaking, the aim of the decoder is to determine the word sequence with the highest likelihood given the lexicon and the acoustic and language models. In practice, however, it is common to search for the most likely HMM state sequence, i.e., the best path through a trellis (the search space) where each node associates an HMM state with given time. Since it is often prohibitive to exhaustively search for the best path, techniques have been developed to reduce the computational load by limiting the search to a small part of the search space. Even for research purposes, where real-time recognition is not needed there is a limit on computing resources (memory and CPU time) above which the development process becomes too costly. The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous Viterbi beam search [108] which relies on a dynamic programming procedure. This basic strategy has been extended to deal with large vocabularies by adding features such as fast match [12], [57], word-dependent phonetic trees [109], forward-backward search [7], N-best rescoring [139], progressive search [107] and simple one-pass dynamic network decoding [112]. An alternative to the frame-synchronous Viterbi beam search is an asynchronous search based on the A* algorithm such as *stack decoding* [11], [124] or the *envelope search* [62].

Dynamic decoding can be combined with efficient pruning techniques in order to obtain a single pass decoder that can provide the answer using all the available information (i.e., that in the models) in a single forward decoding pass over of the speech signal. This kind of decoder, such as the stack decoder [124] or the one-pass frame synchronous dynamic network decoder [112], is very attractive for real-time applications.

Static decoders require much more memory than dynamic decoders when used with long span language models (3-gram or higher order), and as a consequence they are mostly used with smaller language models (usually 2-grams or constrained grammars). It has been recently shown that by proper optimization of a finite-state automaton[1] corresponding to a recognizer HMM network, substantial reduction of the overall network size can be obtained, enabling static decoding with long span LMs [106]. Evidently, the size of the optimized network remains proportional to the LM size.

Many systems under development use multiple pass decoders to reduce the computational requirements if real-time decoding is not an issue [7], [51], [107], [128], [162]. In multipass decoding, additional knowledge sources are progressively used in the decoding process, which allows the complexity of each individual decoding pass to be reduced and often results in a faster overall decoder [110]. For example, a first decoding pass can use a 2-gram language model and simple acoustic models, and later passes will make use of 3-gram and 4-gram language models with more complex acoustic models. This multiple pass paradigm requires a proper interface between passes in order to avoid losing information and engendering search errors. Information is usually transmitted via word lattices or word graphs, or N-best hypotheses. Lattices are graphs where nodes correspond to particular frames and where arcs representing word hypothesis have associated acoustic and language model scores. N-best hypotheses are a list of the most likely word sequences with their respective scores. This multipass approach is not well suited to real-time applications since no hypothesis can be returned until the entire utterance has been processed. However if a small delay is acceptable, then with appropriate synchronization, multipass strategies can be envisioned. Evidently, the first pass used to generate the initial word lattice must be accurate enough to not introduce lattice errors which are unrecoverable with further processing.

E. Adaptation

One of the main challenges in LVCSR is building robust systems that keep high recognition accuracy when testing and training environmental conditions are different. At the acoustic level, two classes of techniques to increase system robustness can be identified: signal processing techniques which attempt to compensate for the mismatch between testing and training by correcting the speech signal to be decoded; and model adaptation techniques which attempt to modify the model parameters to better represent the observed signal. Signal processing based approaches include normalization techniques that remove variabil-

[1] An HMM-based speech recognizer can be seen as a transduction cascade which converts the observed feature vectors to a word string, where to some approximation, each transduction (phone model, word model or language model) can be represented as a finite-state automaton.

ity, thereby increasing the system accuracy under mismatched conditions but often resulting in reduced word accuracy under matched conditions, and compensation techniques which rely on a mismatch model and/or speech model. Model adaptation is a much more powerful approach, especially when the signal processing relies on a speech model. Therefore when computational resources are not an issue, model adaptation is the preferred approach to compensate for mismatches. Model adaptation can be used to reduce the mismatch between test and training conditions or to improve model accuracy based on the observed test data. Adaptation can be of the acoustic models or the language models, or even of the pronunciation lexicon.

Acoustic model adaptation can be used to compensate mismatches of various natures due to new acoustic environments, to new transducers and channels, or to particular speaker characteristics, such as the voice of a non-native speaker. The most commonly used techniques for acoustic model adaptation are parallel model combination (PMC), maximum *a posteriori* (MAP) estimation, and transformation methods such as maximum likelihood linear regression (MLLR). PMC is essentially used to account for environmental mismatch due to additive noise whereas MAP estimation and MLLR are general tools that can be used for speaker adaptation and environmental mismatch.

PMC approximates a noise corrupted model by combining a clean speech model with a noise model [42]. For practical reasons, it is generally assumed that the noise density is Gaussian and that the noise corrupted speech model has the same structure and number of parameters as the clean speech model – typically a continuous density HMM with Gaussian mixture. Various techniques have been proposed to estimate the noisy speech models, including the log-normal approximation approach, the numerical integration approach, and the data driven approach [43]. The log-normal approximation is crude especially for the derivative parameters, and all three approaches require making some approximations to estimate derivative parameters other than first order differences.

MAP estimation can be used to incorporate prior knowledge into the CDHMM training process, where the prior information consists of prior densities of the HMM parameters [54], [89]. In the case of speaker adaptation, MAP estimation may be viewed as a process for adjusting speaker-independent models to form speaker-specific ones based on the available prior information and a small amount of speaker-specific adaptation data. The joint prior density for the parameters in a state is usually assumed to be a product of Normal-Gamma densities for the mean and variance parameters of the Gaussian mixture components and a Dirichlet density for the mixture gain parameters. MAP estimation has the same asymptotic properties as ML estimation but when independent priors are used for different phone models the adaptation rate may be very slow, particularly for large models. It is therefore advantageous to represent correlations between model parameters in the form of joint prior distributions [143], [171].

MLLR is used to estimate a set of transformation matrices for the HMM Gaussian parameters in order to maximize the likelihood of the adaptation data [31], [92], each transform being apply to a subset of the Gaussian pdfs. This adaptation method was originally used for speaker adaptation, but it can equally be applied to environmental mismatch [163]. Since the number of transformation parameters is small, large models can be adapted with small amounts of data. To obtain ML asymptotic properties it is necessary to adjust the number of linear transformations to the amount of available adaptation data. This can be done efficiently by arranging the mixture components into a tree and dynamically defining the regression classes. MLLR adaptation is particularly suited to unsupervised adaptation since the transforms may have a very small number of parameters shared by the different phonetic units and therefore is very robust to recognition errors. In practice only a few regression matrices are used for unsupervised adaptation, usually one or two (corresponding, for example, to speech and non-speech). As a natural extension of this approach, speaker adaptive training (SAT) incorporates supervised MLLR in the SI training procedure and jointly estimate the training speaker MLLR transforms and the HMM parameters [4]. The SAT models which are better suited to MLLR speaker adaptation result in a significant reduction in the error rate by enhancing or boosting the adaptation in particular for supervised adaptation on clean data.

Vocal tract length normalization (VTLN) is another technique which has been proposed to perform some kind of speaker normalization [3]. The approach consists in performing a frequency warping to account for difference in vocal track length, where the appropriate warping factor is chosen from a set of candidate values (typically 13 in the range 0.88 to 1.12 [91]) by maximizing the test data likelihood based on a first decoding pass transcription. Like MLLR adaptation, VTLN can also be applied during the training process to obtain models better suited to decode the normalized test data. VTLN has been shown to give small but significant error rate reduction in particular on telephone conversational speech [151].

Adaptation techniques can evidently also be applied to the language model. In most systems one or more language models are used, but these LMs are usually static, even though the choice of which static model to use can be dependent upon the dialog state, for example. Various approaches have been taken to adapt the language model based on the observed text so far, including the use of a *cache model* [72], [134], a *trigger model* [133], or *topic coherence modeling* [142]. The cache model is based on the idea that words appearing in a dictated document will have an increased probability of appearing again in the same document. For short documents the number of words appearing is small, and as a consequence the benefit is small. The trigger model attempts to overcome this by using observed words to increase the probabilities of other words that often co-occur with the trigger word. In topic coherence modeling, selected keywords in the transcribed speech are used to retrieve articles on similar topics with which sublanguage models are constructed and used to rescore N-best hypotheses. Despite the growing interest in adaptive language models, thus far only minimal improvements have been obtained compared to the use of very large, static *n*-gram models.

III. ENABLING APPLICATIONS

Performance of a speech recognizer is acknowledged to be strongly dependent upon the task, which in turn is linked to the type of user, speaking style, environmental conditions etc. Substantial effort is required to develop a usable system according to the task constraints, even from demonstrated state-of-the-art technology. In adapting a state-of-the-art laboratory speech recognizer for real-world use, all aspects of the speech recognizer must be reconsidered, from signal capture to adaptive acoustic and language models. Given application constraints, standard laboratory development procedures may need to be revised. In this section we summarize some of the issues to be considered from the technology standpoint, such as enhancing the system output with additional annotations (metadata), efficiency and portability across languages and tasks. We do not address human factors or system integration issues which are beyond the scope of this paper.

A. Metadata

More information can be extracted from the audio signal than the simple word string. This additional information, which is useful for higher level processing of the data, can be of a linguistic nature: that is an enhanced transcription (cased text output, punctuation, semantic tags); or of an acoustic nature: speaker turn and identity information, audio type information and confidence measures. For example, although in todays dictation systems the user is required to verbalize all punctuation markers and formatting commands, in the future systems trained on appropriate texts may be able to propose punctuation markers [14].

Semantic tags are another type of linguistic information which can be associated with the transcription. Adding such tags entails applying standard natural language processing (NLP) techniques to an imperfect transcription of the speech rather than to a written text. There are two sources of differences compared with written texts: inherent differences in written and spoken language, and errors due to the automatic transcription process. Example tags can be named entities (names of persons, places, organizations), monetary amounts, dates, times, etc., as well as higher level tags such as topics (e.g., politics, weather report, financial,...). The same HMM-based probabilistic framework can be used to assign tags [103], [157], [164], while also using standard NLP techniques such as tokenization, stemming, stopping, etc. Detailed semantic tagging is often required for dialog tasks where it is common to use task-dependent representations such as semantic frames, with predefined semantic slots and values.

Concerning the acoustic nature of the signal, the same basic modeling techniques can be used to identify other attributes, such as the gender and identity of the speaker [86], and the background acoustic conditions [46], [144]. Such information, when converted to time-aligned markups, can be accessed by search engines. Determining the acoustic structure of the data is the subject of the next subsection.

B. Data partitioning

When transcribing continuous audio streams such as broadcast data, it is advantageous to first partition the data into homogeneous acoustic segments prior to word recognition. Partitioning consists of identifying and removing non-speech segments, and then clustering the speech segments and assigning bandwidth and gender labels to each segment. While it is possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straight-forward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities, and background acoustic conditions. Second, by clustering segments from the same speaker, acoustic model adaptation can be carried out on a per cluster basis, as opposed to on a single segment basis, thus providing more adaptation data. Third, prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. Fourth, by using acoustic models trained on particular acoustic conditions (such as wide-band or telephone band), overall performance can be significantly improved. Finally, eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long), substantially reduces the computation time and simplifies decoding.

Various approaches have been proposed to partition the continuous stream of audio data. Most of these approaches rely on a two step procedure, where the audio stream is first segmented in an attempt to locate acoustic changes (associated with changes in speaker, background or environmental condition, and channel condition). The segmentation procedures can be classified into three approaches: those based on phone decoding [64], [96], [158], distance-based segmentations [81], [146], and methods based on hypothesis testing [23], [159]. The resulting segments are then clustered (usually using Gaussian models), where each cluster is assumed to identify a speaker or more precisely, a speaker in a given acoustic condition. The partitioning approach used in the LIMSI BN transcription system is not based on such a two step procedure, but instead relies on an audio stream mixture model [50]. Each component audio source, representing a speaker in a particular background and channel condition, is in turn modeled by a mixture of Gaussians. The segment boundaries and labels are jointly identified using an iterative maximum likelihood segmentation/clustering procedure using Gaussian mixture models and agglomerative clustering. In contrast to partitioning algorithms that incorporate phoneme recognition, this approach is language independent. (The same models have been used to partition English, French and German data.) The result of the partitioning process is a set of speech segments with speaker, gender and telephone/wide-band labels.

C. Confidence Measures

Confidence measures have been proposed as a way of detecting those hypothesized words that are likely to be erroneous by estimating word and sentence correctness [19], [58], [147], [160], [161]. At the sentence level the goal is to get an estimate of $\Pr(w|x)$ for the hypothesized word string w. One common approach consists of using the posterior $\Pr(w|x,\lambda)$ as an estimate. This assumes that the recognizer models (acoustic model, language model and lexicon designated by λ) are correct and that the decoder does not make any search errors. Further approximations may use simpler acoustic and language models to speed up the computation, for example, the word language model can be replaced by a phone language model [48]. For most LVCSR tasks we are essentially interested by a word level confidence measure, i.e., the goal is to obtain an estimate of $\Pr(w_i|x)$ the posterior probability of the i-th word in the hypothesized word string, or alternatively $\Pr(w_i|x,\lambda)$. An estimate of this latter probability can be efficiently computed by applying the Forward-Backward algorithm to a word graph generated by the speech recognizer [160]. However since this posterior probability relies on incorrect models, it is also common to use additional features such as word and phone durations, speaking rate, and signal-to-noise ratio to better approximate the word posterior probability $\Pr(w_i|x)$. All these predictors can be combined and mapped to the confidence score by using either a logistic regression [58], a generalized additive model [147], or a neural-network [161]. These models are trained on development data by maximizing a confidence score metric such the normalized cross entropy. The proper set of features depends on the particular application.

D. Efficiency

Efficiency of the speech recognizer is not usually a high priority for laboratory systems, where it is typical to develop on loaded (lots of memory and disk space), high powered workstations. The performances of laboratory systems are usually optimized so as to obtain the lowest word error given the training data and the facilities available. However, for commercial products cost is often an important factor which means that the efficiency of the recognizer becomes a higher priority, both in terms of memory and computational requirements, as does the cost of the recognition platform.

Fast decoding techniques are of primary interest, and their requirements influence the choice of model structure and size, and as a result have an impact on the memory needs. For speaker-independent LVCSR based on Gaussian mixture HMM, between 30 and 50% of the recognition time can be spent in computing the HMM state likelihoods, with the remaining time corresponding to the search procedure itself. This is due to the large number of states needed to represent the context-dependent phone models, even when state tying is used. This computation can be reduced either by implementing a fast state likelihood computation which usually requires making some approximations, or by reducing the model size which has the additional advantage of reducing the memory requirements. A widely used technique for speeding up the state likelihood computation is vector quantization of the feature vector space in order to prepare a Gaussian short list for each HMM state and each region of the quantified feature space [17]. With this technique the number of Gaussian likelihoods to be computed during decoding for each input frame and each state can be reduced to a fraction of the number of Gaussians corresponding to the active states with only a small loss in accuracy.

As discussed in section II-D there are many efficient solutions to the search problem, however finding the optimal solution is always a trade-off between the model accuracy and efficient pruning. In general better models have more parameters, and therefore require more computation. However since the models are more accurate, it is often possible to use a tighter pruning level (thus reducing the computational load) without any loss in accuracy. In fact, limitations on the available computational resources can significantly affect the design of the acoustic and language models. For each operating point, the right balance between model complexity and pruning level must be found. Therefore recognizers must be compared at the targeted speed. Aggressive pruning is generally needed to achieve real-time operation for LVCSR tasks on currently available platforms. This inevitably is a source of search errors, and as such, many techniques have been proposed to reduce these search errors and to limit their effect on the recognizer accuracy. One of the most attractive decoding strategies for real-time operation is the one-pass frame-synchronous dynamic network decoder which relies on a phonetic tree organization of the decoding network using LM state conditioned tree copies [6], [109], [112]. The success of such a single pass approach is highly dependent on the use of efficient pruning strategies associated with a language model lookahead [115], [138]. Multipass approaches can also be used successfully for close to real-time operation by chunking the data and running the different pass in parallel with a slight delay.

As explained in section II-A model and state tying are commonly used to improve the model accuracy but optimal tying (from the accuracy point of view) can still result in a very large model with 5 k to 30 k states when large amounts of training data are available. Parameter tying is also powerful technique to reduce the number of parameters, and can be applied to all the levels of the model structure (allophone model, state and Gaussian) [150]. However, more flexibility is available for Gaussian pdf tying in that large model reductions can be obtained without sacrificing too much in terms of system accuracy. This is exemplified by the subspace distribution tying approach [99], [150], which in its most elementary implementation can be seen as a quantization of the model parameters.

Processing time constraints significantly affect the way the acoustic models are selected. For each operating point, the right balance between model complexity and search pruning level must be found. To illustrate this point, Figure 1 plots the word error rate as a function of processing time for 3 sets of acoustic models, which taken together minimize the word error rate over a wide range of processing times (from 0.3xRT to 20xRT) for the LIMSI broadcast news transcription system. (Transcribing such inhomogeneous data requires significantly higher processing power than for speaker adapted dictation systems, due to the lack of control of the recordings and linguistic content, which on average results in lower SNR ratios, a poorer fit of the acoustic and language models to the data, and as a consequence, the need for larger models.) These results on a representative portion of the Hub4-98 data set are obtained on a Compaq XP1000 500 MHz machine with a 3-gram language model, and without acoustic model adaptation. The large model set (350 k Gaussians, 11 k tied states, 30 k phone contexts) provides the best performance/speed ratio for processing times over 5xRT. The 92 k model set (92 k Gaussians, 6 k tied states, 5 k phone contexts) performs better in the range 0.9xRT to 5xRT, whereas a much smaller model set (16 k Gaussians) is needed to go under real-time.

The language model, usually a 3-gram or 4-gram backoff LM in state-of-the-art systems, can have a very large number of parameters (i.e., more than 10 million), and therefore may require prohibitive amounts of memory for commercially viable platforms. One of the attractive properties of n-gram models is the possibility of relying more on the backoff components by increasing the cutoffs on the n-gram counts, thus reducing significantly the LM size. More elaborate n-gram pruning have also

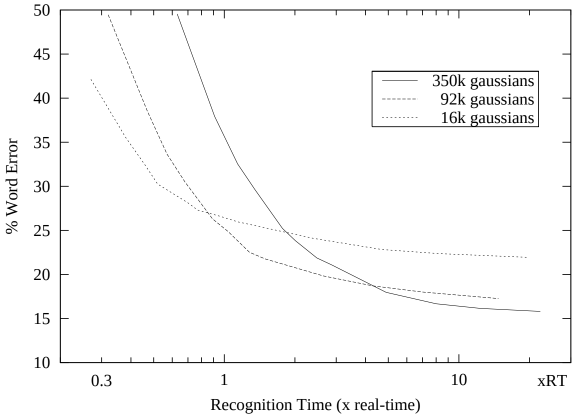

Fig. 1. Word error vs. processing time for three acoustic model sets with 350k, 92k and 16k Gaussians (for broadcast news data).

been proposed [145], [148] to substantially reduce the LM size with negligible loss in accuracy. An alternative approach to limit the memory requirements is to keep most of the LM parameters on the disk, since most n-grams are never used, combined with a cache of the scores for accessed LM states [127].

E. Porting across languages and tasks

Portability is concerned with the porting of technology to new or changing tasks, and/or to other languages. While the same basic speech recognition technology has been successfully used for a variety of tasks and languages, substantial effort is involved to construct the acoustic and language models, and to develop the recognition lexicon. With today's technology, the adaptation of a recognition system to a new task or another language requires the availability of sufficient amounts of transcribed training data. Often, however, the necessary resources are not available and generating them can be long and expensive. Minimizing the required training data (or determining how to optimally acquire such data) remains an outstanding challenge. Yet the performance and development costs largely depend on the available resources and the experience of the system designer.

Acoustic models trained on a sufficiently large and varied corpus (for example a minimum of 10 hours of speech from 100 speakers) appear to be general enough to use as bootstrap models for a new task without task-specific training data if appropriate normalization and compensation techniques are used to reduce differences in the recording conditions (microphone, channel, environmental noise). If speed is an important factor, it can still be interesting to train on task-specific acoustic data to better account for the phonetic coverage of the task.

Language model and lexicon development remain quite task dependent. For some tasks, such as domain-specific dictation, there is a wealth of written texts that can be used for vocabulary selection and language model estimation. For other tasks, in particular for spoken dialog systems, very little (if any) text data may be available, and data collection is an unavoidable development step. Using a recognition system for data collection has been found to be quite effective for such tasks, with successively more accurate systems available as the amount of training data increases [59]. Techniques for adaptation of both the acoustic and language models can greatly improve the performance of a system throughout the development process.

Determining the pronunciation lexicon is often one of the most labor intensive aspects of porting to a new task. Although letter-to-sound conversion programs are available for some languages, these have almost exclusively been developed for speech synthesis purposes and therefore are less appropriate for speech recognition. One of the most common techniques is to make use of a reference lexicon which has been verified (usually both manually and in the context of a system) to serve as a base lexicon. The baseform pronunciations may have been generated using letter-to-sound rules. New words are then added either by using the same letter-to-sound rules, or pronunciation generation tools [82] and often manually corrected. A means of automatically adding new words and pronunciations to the recognition lexicon is crucial for successful deployment of speech technologies.

Although English has been the predominant language for the computer world there has been a large growth in the information available in electronic form (both online and offline) in many of the world's languages. As a result, speech recognition and natural language processing in multiple languages has become a necessity. Building a recognizer for another language is not so different than building a recognizer for a new task, particularly for close languages. Language-dependent system components (such as the phone set, the need for pronunciation alternatives or phonological rules) evidently must be changed. Other language dependent factors are related to the definition and acoustic confusability of the words in the language (such as homophone, monophone, and compound word rates) and the word coverage of a given size recognition vocabulary. Taking into account language specificities can evidently improve recognition performance. For example, tonal languages such as Chinese may benefit from explicit modeling of pitch, which in turn may require modifications to the feature analysis used.

At the lexical level, a given size lexicon will have different coverage across languages and highly inflected languages require a larger lexicon to adequately represent the language. For example, comparing the number of distinct words in comparably sized newspaper text corpora for English, French, German and Italian, the German corpus contains over twice as many distinct words as French, which has more than Italian and English [84].[2] The larger number of distinct words stems mainly from the number and gender agreement in nouns, adjectives and past participles, and the high number of different verb forms. As a consequence, to obtain a lexical coverage of 95% on newspaper texts, an English lexicon need only contain 5000 words, compared to 20,000 for French and Italian, and 65,000 for German.

F. Indexation

One of the main motivations for automatic processing of the audio channels of broadcast data is to serve as a basis for automatic disclosure and indexation for information retrieval (IR) purposes.

[2] The newspaper text corpora compared are the *Wall Street Journal* (English, 37 M words) [5], *Le Monde* (French, 38 M words) [52], *Frankfurter Rundschau* (German, 36 M) [1], and *Il Sole 24 Ore* (Italian, 26 M words) [37], where the total number of words of text material are given in parentheses.

While in traditional IR tasks the result is typically an ordered set of related documents, for spoken document retrieval (SDR) the result is an ordered set of pointers to temporal excerpts [44] or to complete stories if an a priori topic segmentation is available. SDR can support random access to relevant portions of audio or video documents, reducing the time needed to identify recordings in large multimedia databases. The aims of projects like INFORMEDIA [65], THISL [2], and OLIVE [73] are to develop archiving and retrieval systems for broadcast data to enable efficient access to large multimedia digital libraries. OLIVE is also developing tools for cross-lingual access to the archived documents via online query translation.

Automatic text indexation is classically based on document term frequencies, where the terms are obtained after standard text processing, such as text normalization, tokenization, stopping, stemming, query expansion, and named-entity identification [131]. The same techniques have been successfully applied to automatic transcriptions of broadcast news radio and TV documents. Query expansion making use of additional (parallel) sources text data (preferably from the same epoch as the audio data) to locate terms which co-occur with the terms in the original query so as to enrich it, make spoken document retrieval less sensitive to speech recognition errors [74]. In addition topic segmentation and identification are particularly helpful to structure audio streams which, as opposed to text documents, usually have no a priori structure such as story headline and boundaries.

G. Spoken Language Dialog Systems

Spoken language dialog systems (SLDSs) require going beyond transcription to understanding, and incorporate other technologies beyond the focus of this paper, such as dialog management, natural language understanding and generation and speech synthesis. Acoustic signal capture, and integration of speech with other modalities, such as tactile input, are other aspects to be considered. Some of the design issues in developing a speech recognizer for an SLDS are discussed in [15], [27], [32], [35], [48], [49], [59], [61], [63], [125].

Given the nature of interactive dialog, several constraints are placed on the speech recognizer. The primary requirements are for real-time, speaker-independent, recognition of spontaneous speech. What is really meant by real-time speech recognition is that there is a very short delay from when the user finishes speaking and the system responds. An acceptable latency is on the order of 0.5 seconds. This means that speech recognition is being carried out during signal acquisition, in contrast to speech recognizers designed to function in a sentence or segment-based batch processing mode, and requires alternative approaches to cepstral and energy normalization. One straightforward solution is to base the normalization on a window of previously observed frames. Most real-time decoders make use of a single pass search. In order to ensure that a recognition response is given within an acceptable delay, a common solution is to use a dynamic pruning approach, based either on the number of active solutions. If a two-pass search is used, the second pass must be very fast.

It is common practice for the language models of SLDSs to explicitly model filler words and breath noises, as their occurrences are not random, and to use compound words for common word sequences that are subject to strong reduction and coarticulation. Word-class based language models are often used to give a priori information to the speech recognizer and reduce the risk of model inaccuracy due to non-representivity of the training data. Word classes are usually manually specified, but can be automatically derived.

Different approaches have been taken to interface the speech recognizer with the natural language understanding (NLU) component which extracts the meaning of the spoken query. In most systems a bottom up approach is taken, where the output of the recognizer is passed to the NLU component. The recognizer output can be the most probable word sequence, an N-best list of word strings, or a word lattice. In the latter cases, the NLU component can be used to filter the recognizer output. Whether or not there is a need for more than the best word string depends on what information is in the recognition language model and whether more information is available in the NLU. For example, in general the recognizer has limited task domain and world knowledge. So if the best word sequence output by the recognizer is *Wednesday, January thirtieth*, but the thirtieth of January is not a Wednesday, the language understanding component may be able to detect this inconsistency. If the *thirteenth* is both a Wednesday and in an alternative solution, a clarification dialog with the user can be avoided by using this knowledge. The use of dialog context (or dialog state) language models is a way of adding task-specific knowledge in the recognizer [32], [136] and may reduce the need for word graphs or N-best lists.

Most understanding components are based on rules, however some stochastically based systems have been reported [93], [141], [105]. The attraction of statistical methods stems from their success in speech recognition, with human intervention being limited to labeling (or correcting labels). Known disadvantages are that stochastic models require large training corpora in order to reliably estimate model parameters, and the model accuracy is highly dependent upon the representivity of the training data. Also, generalizations that can be made relatively easily by humans may not be automatically learned.

A confidence measure can be associated with each word in the output, and uncertain words can be rejected by the recognizer or the higher level understanding components, or confirmed via a confirmation subdialog. Rejection has strong implications for the interaction with the user (there is a risk of annoying the user by asking for a repetition) and on average leads to longer dialogs. However, this may be preferable to making an error, and may be more successful in the long run.

IV. APPLICATIONS

This section addresses three main classes of applications based on LVCSR technology, and provides some specific examples taken from our experience at LIMSI. We do not attempt to provide an exhaustive survey of available systems, but rather aim to highlight some application areas of recent attention in the community.

Dictation is the most obvious application of automatic speech recognition technology, as is evidenced by long history of research and product development and the availability of low-cost, off-the-shelf systems for a variety of platforms and languages. Perhaps the most notable characteristic of this task is that the speech data is being produced with the explicit goal of being transcribed by a machine.

The second application area goes beyond dictation to the transcription and indexation of more general audio data, such as radio and television broadcasts, or meetings and teleconferences, and any kind of audio data mining. Several characteristics of this type of audio data can be noted. First, it can be considered "found" data in that it is produced for other reasons, and it is only a secondary benefit to be able to automatically structure the data for other uses. Second, the data consists of a continuous audio stream, where there are multiple speaker turns (maybe overlapping), and there is no a priori segmentation into sentences. Third, the signal capture and background environment can be only more or less controlled. The earliest work in this area that we are aware of is the NSF INFORMEDIA project [65] under the Digital Libraries News-on-Demand action line. A special section of the Communications of the ACM was recently devoted to this topic [102].

The third application class is that of dialog systems. For the most part such systems aim to enable vocal access to stored information. While there has recently been an emergence of dialog systems on the market, the dialog capability of these systems is usually more constraining than laboratory prototypes. We do not address the class of small vocabulary ASR systems as are starting to be seen in telephony applications, such as automated operator assistance or call routing where the keypad menu selection is replaced by vocal commands.

A. Dictation

The first commercially available products based on large vocabulary automatic speech recognition technology were for the dictation task, and today a variety of software-only continuous speech dictation systems are available for the general public. Two main types of dictation tasks can be considered: general dictation and dictation in specific domains. The first task concerns dictation of letters or email, and various other texts. Dictation for specific domains has mainly addressed the legal and medical fields and subspecialties, where there has been a long tradition of dictation services. Another dictation task, that of aids for language learning, is not considered here. While from the technological viewpoint, dictation is usually thought of as the "simple" transformation from speech to text, this view overlooks a variety of formatting and integration issues which are important for products.

The speech data input for a single dictation session is usually from a single speaker and has a restricted linguistic content. The data is close to read speech, and may even be produced from a handwritten manuscript. Even if the text is not written in advance, the speech can be considered "prepared" in that the speaker has planned what text to say. The word stream is also quite close to the written form, since the result will conform to the rules of the written language and not those of spoken language. The microphone can be selected by the system developer, and is usually a close-talking headset mounted microphone. Most systems have a push-to-talk control (or an equivalent sleep/wake-up command) to let the recognizer know when it should be transcribing.

The first commercially available dictation systems were speaker-dependent, requiring an initialization session in which speaker-specific training data was obtained, and for the most part, recognized isolated words. IWR mode provided two main advantages: simplification of the decoding process and of the means for error correction. Today most systems make use of speaker-independent acoustic models that are adapted on-line to the new user with little or no explicit enrollment. Efficient model adaptation techniques (as discussed in Section II-E) are used to minimize the need for speaker-specific data thus vastly improving the perceived system usability.

An advantage of the dictation task from the developers viewpoint is that it is relatively simple to evaluate the core technology by comparing the system hypothesis to a reference word transcription. As such, dictation has served as a baseline performance measure in LVCSR, most notably in the benchmark tests sponsored by the US DARPA programs and coordinated by NIST (National Institute for Science and Technology). This close relation between system development and evaluation, which has been referred to as "assessment driven technology development" had led to larger performance improvements despite increasingly difficult tasks. The commonly used error metric is the "word error" rate defined as: *%word error* = *%substitutions* + *%insertions* + *%deletions*. For the DARPA benchmarks, a case-insensitive text form has always been used to measure the word error rate. For read speech tasks, the state-of-the-art in speaker-independent continuous speech recognition in 1995/1996 [119], [120] is exemplified by the benchmark tests on North American Business News task. The acoustic training data was comprised of about 160 h of read newspaper texts from several hundred speakers and the language model training material was comprised of 400 M words of newspaper texts, from a variety of sources. On test data recorded with a close-talking microphone with an SNR of about 30 dB, word error rates around 7% were obtained using a 65 k word vocabulary.[3] The same read speech recorded with a table-top microphone in a computer room/office environment (noise level 55 dBA, SNR about 15 dB), resulted in a word error of about 14% with noise compensation. Without noise compensation the word error rates of systems trained on only clean speech data is over 50%. The word error for read newspaper texts recorded over long distance telephone lines was over 20%. Spontaneous dictation of business and financial news was addressed by asking subjects with experience in journalism to read about a subject and then dictate a text. The journalists were not allowed to read from a draft, but were allowed to reject ill-formed sentences [80]. The word error on this data was about 14%. Another task addressed speech recognition of non-native talkers. With a set of 40 adaptation sentences, speaker adaptation reduced the word error rate by 2 (from 21% to 11%). Although not an official benchmark result, comparable word error reductions have been obtained for native speakers on other tasks.

While the results given here are for American English, somewhat comparable results have been reported by various sites for other languages. The LRE SQALE (Speech recognizer Quality Assessment for Linguistic Engineering) project [167], which aimed to assess language-dependent issues in multilingual rec-

[3] With the exception of the telephone recordings, the speakers were allowed to repeat a recording if s/he noticed an error or were not satisfied.

ognizer evaluation, demonstrated that the same recognition technology and evaluation methodology used for American English could be successfully applied to British English, French and German.

B. Audio Indexing

Automatic speech recognition is a key technology for audio and video indexing, for data such as radio and television broadcasts. The transcription and indexation of speech recorded at meetings, workshops and teleconferences has many similarities to broadcast data. The transcription of such data presents new challenges as the signal is one continuous audio stream that contains segments of different acoustic and linguistic natures.

The characteristics of this type of data are quite different those of data input to most speech recognizers in the past. Up until the last few years, speech recognizers have been confronted primarily with read or prepared speech, as in dictation tasks where the speech data is produced with the purpose of being transcribed by the machine, or with limited domain spontaneous speech in more-or-less system driven dialog systems. In all cases, the user can adapt his/her language to improve the recognition performance, which can be crucial for some applications. An interesting aspect of the broadcast news domain is that, at least for what concerns major news events, similar topics are simultaneously covered in different emissions and in different countries and languages. Automatic processing carried out on contemporaneous data sources in different languages can serve for multilingual indexation and retrieval. Multilinguality is thus of particular interest for media watch applications, where news may first break in another country or language.

Radio and television broadcast shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic natures. The signal may be of studio quality or may have been transmitted over a telephone or other noisy channel (i.e., corrupted by additive noise and nonlinear distortions), or can contain speech over music or pure music segments. Gradual transitions between segments occur when there is background music or noise with changing volume, and abrupt changes are common when there is a switch between speakers in different locations. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc. Speech from the same speaker may occur in different parts of the broadcast, and with different background noise conditions. The linguistic style ranges from prepared speech to spontaneous speech. Acoustic and language modeling must accurately account for this varied data.

Two principle types of problems are encountered in automatically transcribing audio data streams: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Noise robustness is also needed in order to achieve acceptable performance levels. In order to be robust with respect to the varied acoustic conditions, the acoustic models are typically trained on large corpora (several tens of hours to over a hundred hours) containing all data types. Band-limited acoustic models are often used for segments labeled as telephone speech.

The linguistic models are similarly trained on large text corpora from various sources with different linguistic properties, such as newspaper and newswire texts, Internet data, commercial transcriptions and detailed transcriptions of acoustic data. For example, the LIMSI American English language models result from the interpolation of 3 language models trained on different sources: 200 million words of commercial broadcast news transcriptions; 350 million words of North American Business newspapers and Associated Press Wordstream texts; and 1.6 million words corresponding to the transcriptions of the broadcast news acoustic training data. The importance of the accurate transcriptions can be seen in that the interpolation coefficient of this data is .25, despite the limited amount available. In fact, there is only a slight performance degradation (under 2% relative) if only the commercial transcripts and acoustic data transcripts are used for LM training.

Most of todays state-of-the-art systems for transcription of broadcast data employ the techniques described in Section II, such as PLP features with cepstral mean and variance normalization, VTLN, unsupervised MLLR, decision tree state tying, gender- and bandwidth-specific acoustic models. The recognition vocabulary contains 65,000 or more words, with a lexical coverage over 99% on the American English broadcast news data. Given the spontaneous nature of parts of the audio data, it is important to explicitly model filler words and breath noise [46], which are less common in dictation.

Word recognition is generally performed in two or more decoding passes. The first pass is used to generate an initial word hypothesis, which is used for unsupervised cluster-based acoustic model adaptation. This adaptation, which aims to reduce the mismatch between the models and the data, is needed for generating accurate word hypotheses. When multiple decoding passes are carried out, information is usually transmitted via word graphs or lattices.

Over the last 4 years tremendous progress has been made on transcription of broadcast data [121], [122], [123]. State-of-the-art transcription systems achieve word error rates around 20% on unrestricted broadcast news data, with a word error of about 15% obtained on the recent NIST test sets which were selected to include of higher proportions of studio and announcer data [39]. Transcription performance varies quite a bit across the data types. The average word error rate reported on prepared, announcer speech was about 8% in the DARPA'98 benchmark data and under 2% for some speakers. Performance decreased substantially for spontaneous portions (average word error 15%), degraded acoustic conditions (average word error 16%), or speech from non-native speakers (over 25%).

The transcription of broadcast data has also been a recent focus of research efforts in several other languages, including French, German, Italian, Japanese, Mandarin and Spanish [18], [73], [77], [114], [123] using the same technology. The reported error for these languages are somewhat higher than for American English which can be at least partially attributed to the smaller amounts of training data available in other languages, in particular to the difficulty of obtaining commercial transcripts for language model estimation. For example, in the context of the LE-OLIVE project, we have developed transcription systems for French and German, with word error rates around 30%

higher than the best reported results for American English.

The same technology can be applied to other problems, such as the transcription of meetings and conferences, or telephone conversations (help lines, call centers). Each of these tasks poses a set of specific problems with regard to signal capture (single or multiple channels), speaker population, speaking style and linguistic content, etc. The closest task for which speech recognition results are publicly available is the DARPA Hub5 conversational speech recognition task using the Switchboard [60] and multilingual Callhome (Spanish, Arabic, Mandarin, Japanese, German) corpora. The word rates reported for this data, on the order of 30-40% [168], are substantially higher than those for broadcast news. The Callhome data is particularly challenging to transcribe as the conversations are between two people that know each other, and speak in a familiar manner about subjects of common interest. In addition there are varied acoustic conditions with respect to the background environment and the telephone channel.

As part of the SDR'99 TREC-8 evaluation 500 hours of unpartitioned, unrestricted American English broadcast data were indexed using both state-of-the-art speech recognizer outputs and manually generated closed captioning [45], [155]. The average word error measured on a representative 10 hour subset of this data was around 20% for state-of-the-art systems [45]. It is important to note that not all errors are important for information retrieval, particularly since most information retrieval systems first normalize word forms (stemming). Only small differences in information retrieval performance were observed for automatic and manual transcriptions when the story boundaries are known, indicating that the transcription quality may not be a limiting factor on IR performance for current IR techniques.

C. Spoken Language Dialog

There are many potential services that are based on spoken language dialog systems. The simplest, which are starting already to enter the marketplace, are quite similar to DTMF-based voice response systems, with little requirements for natural language understanding and with relatively constrained dialogs. One example, is call routing services which range from relative small vocabulary (100-500 words) tasks, such as automatic standards in small companies, to several thousand words for standards at large organizations or on-line help services. One of the most explored application domains is that of travel information services, but other areas have also been of interest such as stock quotations, weather information, names, addresses and telephone numbers, used car sales, insurance policies and general tourist information, to mention a few.

In order to enable user-friendly interaction with a machine, it is necessary to be able to recognize naturally spoken spontaneous utterances. It cannot be assumed that the user will be familiar with the system (or with speaking to computers), and in general a user can be expected to interact only briefly with the machine, so there is very little data available for model adaptation. In certain targeted applications it may be possible to have a known user group, in which case this additional information can be used to improve the overall transaction performance.

In contrast to a dictation application where it is relatively straight-forward to select a recognition vocabulary from large written corpora, for specific tasks, there usually are no application-specific training data (acoustic or textual) available. It is therefore necessary to collect application-specific data, which is needed for accurate modeling at different levels (acoustic, lexical, syntactic and semantic). This data collection represents a significant portion of the SLDS development effort [87]. Acquiring sufficient amounts of LM training data is more challenging than obtaining acoustic data. With 10 k queries relatively robust acoustic models can be trained, but this number of queries will typically contain fewer than 100 k words, which may not be sufficient for word list development or for training *n*-gram language models, and are unlikely to yield a complete coverage of the task.

Two broad classes of applications are considered: telephone-based services and multimedia interfaces. Telephone services are a natural area for spoken dialog systems as the only means of interaction with the machine are via voice[4] and have thus been the focus of many development efforts. Since all interaction with the caller is by speech, dialog design and response generation are very important aspects of the system, particularly in the context of natural, mixed-initiative systems where the user is free to change the direction of the dialog at essentially any point in time. Therefore careful consideration must be given to the content and formulation of clear and concise system responses.

Information kiosks and multimedia web interfaces are spreading in availability, providing different ranges of services, such as automated ticketing, orientation information, and general tourist services. Audio output (both sound and speech) can be used to direct the users attention or to provide information. For most multimedia interfaces, the input modalities are limited to a touch screen and a keyboard, however there is increasing interest in speech as an alternative input modality.

Although these 2 application classes share many commonalities, there are important differences that should be pointed out. The main differences concern dialog strategies and signal capture. By necessity, dialog plays a much more important role in telephone-based services, where in general multiple caller-system turns are required to obtain a satisfactory response. For example, it is preferable to ask the caller to provide additional constraints to limit the possible solutions, then to simply read off a long list of possible solutions satisfying a request. With a multimedia interface it can be more efficient to display all possibilities on the screen, letting the user select amongst them.

Signal capture considerations are also quite different. Telephone signal has reduced bandwidth, and may be affected by channel distortions and varyied handset characteristics. For multimedia interfaces a wide-band signal is available, but the microphone is generally far from the talker's mouth. In order to account for different heights and positions of the expected user population, it may be desirable to use multiple microphones [48]. One obvious solution is to use a handset to control the microphone position, but this has the disadvantage of reducing the user's freedom to use other input modalities. Noisy background acoustic conditions are to be expected for multimedia interfaces located in public places. Background noise can

[4] There are still large populations that do not have touch tone access, and the ergonomics of keypad input with the popular telephone design of keys on the handset is not evident!

evidently also be a problem for telephone services if the call is made from a noisy place.

For both types of applications the capability of the user to interrupt the machine is often considered as crucial for usability. (There may of course be dialog contexts where it is desirable to disable barge-in to ensure that the caller hears the entire message.) For the telephone application echo cancelation must be used to remove the echo of the known synthetic speech in order to recognize what is said by the caller. Evidently barge-in which is based on the recognizer output, and not just speech detection, is more efficient and less prone to errors. Simple energy based techniques can be triggered by spurious noises, which can be generated by the user (coughing, throat clearing, touching the microphone) or externally (tapping, door slam, paper rustling). Barge-in with multimedia interfaces requires acoustic echo cancelation, which is a difficult task as the user is generally in the acoustic field and any movement changes the filter characteristics.

Using speech technology to improve the usability of kiosks was addressed in the ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) project, aimed at providing access to rail travel information via a kiosk located in a Parisian train station [47]. The MASK kiosk allows both vocal and tactile input. Early in the project a study was carried out of assess ways if combining input modalities, and it was found that even when given the opportunity, subjects did not mix input modes within a single utterance [94]. In fact, subjects typically preferred either speech or touch, and only switched modes if they experienced problems.[5] Experiments were also carried out to assess the user acceptance of touch-to-talk, which greatly simplifies the speech detection problem, and avoids processing queries not directed to the system. It turned out that the subjects found touch-to-talk to be reassuring as they knew when the system was listening. (Evidently the touch-to-talk is only used to get a rough estimate of the query endpoints as users inevitably speak earlier or later than they touch.)

The most widely known efforts in evaluation of SLDSs are the DARPA ATIS task [66], [98], [126], the German national Verbmobil project [156] and the EC Language Engineering projects [100], [101]. Some recent European activities include the ESPRIT MASK and the LE RAILTEL, MAIS and ARISE projects [16], [113]. The word error rates of the best systems reported in the DARPA ATIS benchmark tests [118], [119] are under 5% for high quality laboratory data, and the spoken language system (SLS) understanding error based on the spoken input is not much larger than the NL understanding error obtained using the orthographic transcription of the query.

More generally, a wide range of word error rates have been reported for the speech recognition components of a spoken dialog systems, ranging from under 5% for simple travel information tasks using close-talking microphones to over 25% for telephone-based information retrieval systems. It is quite difficult to compare results across systems and tasks as different transcription conventions and text normalizations are often used. Also these numbers can be misleading as the word error measures all differences between the exact orthographic form of the query and the recognizer output, and some of recognition errors (such as gender or plurals) are not important for understanding.

While there are commonly used measures and methodologies for evaluating speech recognizers, the evaluation of spoken dialog systems is considerably more complicated due to the interactive nature and the human perception of the performance [15], [27], [98]. It is therefore important to assess not only the individual system components, but the overall system performance using objective and subjective measures [88], [104]. For example, in addition to the commonly used word error rate, it can be enlightening to measure the error on words that are important for the task. Some objective measures of the global system performance include the success rate, the average/maximum/minimum number of turns, the total/waiting time, the number of repetitions. In the case of multimodal systems, the effectiveness of speech can be compared with other modalities, such as touch screen or keypad for input and a visual display for output.

D. Challenges and Perspectives

Despite the numerous advances made over the last decade, speech recognition is far from a solved problem, as evidenced by the large gap between machine and human performance [29], [36], [95], [152]. The performance difference is a factor of 5 to 10, depending upon the transcription task and test conditions. To reduce this difference further improvements are needed in the modeling techniques at all levels: acoustic, lexical and linguistic (syntactic and semantic).

It is well acknowledged that for laboratory systems (to the best of our knowledge no performance measures are available for commercial dictation systems) there can be a huge performance difference, such as a factor of 20 or more in the word error rates for the best (1-2%) and worst speakers (25-30%). This can be attributed to a variety of factors [38] mainly, the speaking style and speaking rate. For moderate speaking rates (120-160 words per minute), there is no strong correlation between speaking rate and word error rate, however, for speaking rates over 180 words per minute, the word error rate increases significantly [119]. Acoustic model adaptation can partially reduce this difference, but requires several minutes of data to be efficient, which limits its use. Faster adaptation techniques which can better account for the correlation between the parameters of the model are therefore needed. Reducing this difference may also require adaptive pronunciation models, which can predict pronunciation variants based on the observed pronunciations for the given speaker. A person who pronounces a word in a given manner is likely to pronounce similar words in a similar way. Similarly, at the cross-word level, different speakers make use of different phonological rules. Although these rules are usually systematic, no systems that we know of are able to make use of this consistency.

Even with an average word error rate of 5% for speaker adapted dictation systems, the user must correct one out of twenty words, which is a costly process. An analysis of real users' experience with dictation, comparing the efficiency of

[5] An important difference in dialog strategies is offered by the two input modes. Tactile input is based on a menu driven dialog, where the user must input specific information in order to move on to the next step. Vocal input allows a real mixed-initiative dialog between the user and the system, where the user can guide the interaction or be guided by the system via the help messages.

dictation with typing is given in [75].

One class of future potential products based on dictation technology are telephone services offering the ability to dictate a letter, fax or email message. However, before such applications can become widespread, performance will need to be improved. Extrapolating from the results given above for spontaneous journalist dictation and for read telephone speech, expected word error rates for spontaneous dictation over the telephone are likely to be over 30%. Distributed speech recognition, where acoustic parameterization is carried out on the local handset or webphone, and the coded parameters transmitted to a central server for recognition, may help solve this problem by eliminating the variability due to the telephone channel.

Concerning language modeling, to date techniques for longer term agreement have resulted in only minimal improvements. They should however be useful for accurate transcription of highly inflected languages where *3*-grams are clearly not the optimal solution.

Keeping the language model up-to-date is a challenge for broadcast news transcription due to the the fast, changing nature of news. New topics appear suddenly, and remain popular for quite variable length time periods. One of the most difficult problems is to be able to recognize previously unseen or rare proper names. Fortunately other sources of contemporary data are available to help keep the system up-to-date, such as written documents from newspapers and newswires, many now available on the Internet, which can be used by the transcription system to continually update its lexicon and language model. This is not a trivial problem since producing phonetic transcriptions of new words such as proper names (in particular for foreign names which are quite common in broadcast data) must rely on some acoustic evidence, since the pronunciation of foreign words can be quite variable depending upon the talker's knowledge of the foreign language.

Developing systems for many languages at reasonable cost is a problem that may require less supervised training procedures. Some very promising work has been recently reported by [77] using untranscribed training data for acoustic model estimation. An initial system is developed using a small amount of training data (10 hours). This system is then used to transcribe a second set of data, and models are reestimated. The new models are then used to transcribe more data, and the cycle is reiterated.

In our view, the main challenge of spoken language dialog systems is to provide a natural, user-friendly interface with the computer, allowing easy access to the stored information. The user should be free to ask any question or to provide any information at any point in time, but the system should help the user if the user appears to be in difficulty. We have observed that some speakers had serious difficulty in interacting with the ARISE system, and suspect that there is a class of users that will experience similar difficulties with any such system. How large a percentage of the targeted user population falls into this category of user is unknown. Even for deployed systems, evaluation is carried out on the calls that are received, by default eliminating people that have called the system only once and never called back. Speech recognition for SLDSs is complicated by the fact that speaker-independent modeling is a necessity, as the total amount of speech during any interaction is small so that it is difficult to take advantage of model adaptation. As discussed above, this results in a wide range in recognition errors across speakers, and in particular for speech from non-native speakers, for whom the word error can be twice as high as for native ones [59]. Also, in order to improve speech recognition performance on spontaneous speech it may be interesting to question the basic units used for acoustic modeling, as units other than context-dependent phones may prove to better capture the large amount of phonological variants. For language modeling a similar question can be posed regarding how to better model contractions and sloppy articulation resulting in word deletions.

Task independence is another outstanding challenge, particularly concerning the language models. If sufficient acoustic training data is available, it is possible to estimate acoustic models that work pretty well for a variety of tasks. This is not the case for language models, where domain coverage is critical. Constructing corpora that are representative, complete, and yet at the same time not too big, remains an open research area in spite of our collective experience.

Although it is generally advocated that speech can provide a more natural interface with the computer than a keyboard or a mouse, few studies have addressed multimodal interaction using speech. User trials of the MASK kiosk [83] carried out with over 200 subjects demonstrated that for this task multimodality is more efficient (faster and easier) than unimodality as some actions are better carried out by voice and others by touch. These studies also showed that subjects performed their tasks more efficiently as they became familiarized with the MASK system, learning to exploit the vocal input and benefiting from the multiple modalities. Audio-visual speech recognition [132] is a promising research direction to improve the usability of multimodal kiosks.

V. CONCLUSIONS

The last decade has witnessed significant advances in speech recognition technology. The move from processing of prepared speech separated in sentences to continuous inhomogeneous audio streams is one of these major advances. This capability has been enabled by advances in techniques for robust feature extraction, acoustic modeling with effective parameter sharing, unsupervised adaptation to speaker and environmental condition, efficient dynamic network decoding, and audio stream partitioning algorithms.

Even though substantial progress has been made, machine performance is still a long way behind human performance. Transcription of spontaneous speech remains quite challenging in part due to the large variety in speaking style and fluency. While it is clear that all our models could use improvement, it is not clear which of acoustic modeling, language modeling or the phonetic lexicon is the weakest link. In fact, we have difficulties in modeling distant dependencies at all levels.

Ongoing research is addressing issues such as low cost system development, lightly supervised training, faster adaptation techniques, learnable pronunciation lexicons, language model adaptation, topic detection and labeling, and metadata annotation.

A wide range of potential applications can be envisioned based on current technology, particularly in the area of automatic indexation of broadcast data, where automated processing is a necessity to keep up with the flow of information. This is an exciting research area, in that there are many outstanding issues to be addressed to improve the transcription accuracy on this varied data, and at the same time there are near-term applications which can be successfully built upon this technology.

REFERENCES

[1] ACL-ECI CDROM, distributed by Elsnet and LDC.

[2] D. Abberley, D. Kirby, S. Renals and T. Robinson , "The THISL Broadcast News Retrieval System," *Proc. ESCA ETRW on Accessing Information in Spoken Audio*, pp. 14-19, Cambridge, U.K., April 1999.

[3] A. Andreoum T. Kamm and J. Cohen, "Experiments in Vocal Tract Normalisation", *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[4] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A Compact Model for Speaker Adaptation Training", *Proc. ICSLP'96*, pp. 1137-1140, Philadelphia, PA, October 1996.

[5] "CSR corpus. Language model training data," *NIST Speech Disc 22-1 and 22-2*, Produced by LDC, August 1994.

[6] X. Aubert, "One Pass Cross Word Decoding for Large Vocabularies Based on a Lexical Tree Search Organization," *Proc. ESCA Eurospeech'99*, **4**, pp. 1559-1562, Budapest, Hungary, September 1999.

[7] S. Austin, R. Schwartz and P. Placeway, "The Forward-Backward Search Strategy for Real-Time Speech Recognition," *Proc. IEEE ICASSP-91* pp. 697-700, Toronto, Canada, May 1991.

[8] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer and H.F. Silverman,"Preliminary results on the performance of a system for the automatic recognition of continuous speech," *Proc. IEEE ICASSP-76*, Philadelphia, PA, April 1976.

[9] L.R. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan and S. Roukos, "The IBM Large Vocabulary Continuous Speech Recognizer for the ARPA NAB News Task," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 121-126, Austin, TX, January 1995.

[10] L.R. Bahl, P. Brown, P. de Souza, R.L. Mercer and M. Picheny, "Acoustic Markov Models used in the Tangora Speech Recognition System," *Proc. IEEE ICASSP-88* **1**, pp. 497-500, New York, NY, April 1988.

[11] L.R. Bahl, F. Jelinek and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis & Machine Intelligence*, **PAMI-5**(2), pp. 179-190, March 1983.

[12] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo and M. Picheny, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *Proc. IEEE ICASSP-92*, CA, **1**, pp. 17-21, San Francisco, CA, March 1992.

[13] J. Baker, J. Baker, P. Bamberg, K. Bishop, L. Gillick, V. Helman, Z. Huang, Y. Ito, S. Lowe, B. Peskin, R. Roth and F. Scattone, "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems," *Proc.* DARPA *Speech & Natural Language Workshop*, pp. 387-392, Harriman, NY, February 1992.

[14] D. Beeferman, A. Berger and J. Lafferty, "Cyberpunc: A Lightweight Punctuation Annotation System for Speech," *Proc. IEEE ICASSP-98*, **2**, pp. 689-692, Seattle, WA, May 1998.

[15] N.O. Bernsen, L. Dybkjaer and U. Heid, "Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems," *Proc. ESCA Eurospeech'99*, pp. 1147-1150, Budapest, Hungary, September 1999.

[16] M. Blasband "Speech Recognition in Practice: The ARISE Project (Automatic Railway Information System for Europe)," La lettre de l'IA numéros 134-135-136, *Proc. NIMES'98*, pp. 207-210, Nîmes, France, June 1998.

[17] E. Bocchieri, "Vector quantization for efficient computation of continuous density likelihoods," *Proc. IEEE ICASSP-93*, **2**, pp. 692-695, Minneapolis, MN, May 1993.

[18] F. Brugnara, M. Cettolo, M. Federico and D. Giuliani, "A Baseline for the Transcription of Italian Broadcast News," *Proc. IEEE ICASSP-00*, Istanbul, Turkey, June 2000.

[19] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition", *Proc. ESCA Eurospeech'97*, pp. 815-818, Rhodes, Greece, September 1997.

[20] L. Chase, R. Rosenberg, A. Hauptmann, M. Ravishankar, E. Thayer, P. Placeway, R. Weide and C. Lu, "Improvements in Language, Lexical and Phonetic Modeling in Sphinx-II," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 60-65, Austin, TX, January 1995.

[21] F. Chen, "Identification of contextual factors for pronunciations networks," *Proc. IEEE ICASSP-90*, pp. 753-756, Albuquerque, NM, April 1990.

[22] S.F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer, Speech & Language*, **13**(4), pp. 359-394, October 1999.

[23] S.S. Chen, P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 127-132, Landsdowne, VA, February 1998.

[24] Y.L. Chow, R. Schwartz, S. Roukos, O. Kimball, P. Price, F. Kubala, M.O. Dunham, M. Krasner and J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", *Proc. IEEE ICASSP-86*, **3**, pp. 1593-1596, Tokyo, Japan, April 1986.

[25] P. Clarkson and R. Rosenfeld, "Statistical Language modelling using CMU-Cambridge Toolkit," *Proc. ESCA EuroSpeech'97*, pp. 2707-2710, Rhodes, Greece, September 1997.

[26] M. Cohen, *Phonological Structures for Speech Recognition*, PhD Thesis, U. Ca. Berkeley, 1989.

[27] G. Dafydd, R. Moore and R. Winski, Eds., *Handbook of standards and resources for spoken language systems*, Mouton de Gruyter. Berlin, New York. 1997.

[28] S. Davis and P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, & Signal Processing*, **28**(4), pp. 357-366, *month* 1980.

[29] N. Deshmukh, A. Ganapathiraju, R.J. Duncan and J. Picone, "Human Speech Recognition Performance on the 1995 CSR Hub-3 Corpus" *Proc. ARPA Speech Recognition Workshop*, pp. 129-134, Harriman, NY, February 1996.

[30] V. Digalakis and H. Murveit, "Genones: Optimization the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer," *Proc. IEEE ICASSP-94*, **1**, pp. 537-540, Adelaide, Australia, April 1994.

[31] V. Digalakis, D. Rtichev and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", *IEEE Trans. on Speech & Audio*, **3**(5), 357-366, September 1995.

[32] E.W. Drenth and B. Rüber, "Context-dependent probability adaptation in speech understanding," *Computer Speech & Language*, **11**(3), pp. 225-252, July 1997.

[33] J. Dreyfus-Graf, "Sonograph and Sound Mechanics," *J. Acoust. Soc. America*, **22**, pp. 731, *month* 1949.

[34] H. Dudley and S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," *J. Acoust. Soc. America*, **30**, pp. 721, *month* 1958.

[35] L. Dybkjaer, N.O. Bernsen, R. Carlson, L. Chase, N. Dahlbäck, K. Failenschmid, U. Heid, P. Heisterkamp, A. Jönson, H. Kamp, I. Karlsson, J. v.Kuppevelt, L. Lamel, P. Paroubek and D. Williams, "The DISC Approach to Spoken Language Dialog Systems Development and Evaluation," *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pp. 185-189, Granada, Spain, May 1998.

[36] W.J. Ebel and J. Picone, "Human Speech Recognition Performance on the 1994 CSR Spoke 10 Corpus" *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 53-59, Austin, TX, January 1995.

[37] M. Federico, M. Cettolo, F. Brugnara and G. Antoniol, "Language Modeling for Efficient Beam-Search," *Computer Speech & Language*, **9**(4), 353-379, October 1995.

[38] W.M. Fisher, "Factors Affecting Recognition Error Rate," *Proc. ARPA Speech Recognition Workshop*, pp. 47-52, Harriman, NY, February 1996.

[39] W.M. Fisher, W.S. Liggett, A. Le, J.G. Fiscus and D.S. Pallett, "Data Selection for Broadcast News CSR Evaluations," *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 12-15, Landsdowne, VA, February 1998.

[40] E. Fosler, M. Weintraub, S. Wegmann, Y.-H. Kao, S. Khudanpur, C. Galles and M. Saraclar, "Automatic learning of word pronunciation from data," *Proc. ICSLP'96*, **Addendum**, pp. 28-29, Philadelphia, PA, October 1996.

[41] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. on Acoustics, Speech & Signal Processing*, **ASSP-29**, pp. 342-350, *month* 1981.

[42] M.J.F. Gales and S.J. Young, "An improved approach to hidden Markov model decomposition of speech and noise," *Proc. IEEE ICASSP-92*, pp. 233-236, San Francisco, CA, March 1992.

[43] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *Computer Speech & Language*, **9**(4), pp. 289-307, October 1995.

[44] J.S. Garofolo, E.M. Voorhees, C.G.P. Auzanne, V.M. Stanford and B.A. Lund, "1998 TREC-7 Spoken Document Retrieval Track Overview and Results", *Proc. 7th Text Retrieval Conference TREC-7*, NIST Special Publication 500-242, pp. 79-90, Gaithersburg, MD, November 1998.

[45] J.S. Garofolo, C.G.P. Auzanne, E.M. Voorhees and W.M. Fisher, "1999 TREC-8 Spoken Document Retrieval Track Overview and Results", *Notebook of the 8th Text Retrieval Conference TREC-8*, Gaithersburg, MD, November 1999.

[46] J.L. Gauvain, G. Adda, L. Lamel and M. Adda-Decker, "Transcribing

Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, pp. 56-63, Chantilly, VA, February 1997.

[47] J.L. Gauvain, S. Bennacef, L. Devillers, L. Lamel and R. Rosset, "Spoken Language component of the MASK Kiosk" in K. Varghese, S. Pfleger(Eds.) "Human Comfort and security of information systems", Springer-Verlag, 1997. Also in *Proc. Human Comfort and Security Workshop*, Brussels, Belguim, October 1995.

[48] J.L. Gauvain, J.J. Gangolf, L. Lamel, "Speech Recognition for an Information Kiosk," *Proc. ICSLP'96*, pp. 849–852, Philadelphia, PA, October 1996.

[49] J.L. Gauvain and L. Lamel, "Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications," *Institute of Electronics, Information and Communication Engineers*, J79-D-II:2005–2021, December 1996.

[50] J.L. Gauvain, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, **5**, pp. 1335-1338, Sydney, Australia, December 1998.

[51] J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker, "The LIMSI Nov93 WSJ System," *Proc. ARPA Spoken Language Technology Workshop*, pp. 125-128, Princeton, NJ, March 1994.

[52] J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, October 1994.

[53] J.L. Gauvain, L.F. Lamel and M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *Proc. IEEE ICASSP-95*, pp. 65-68, Detroit, MI, May 1995.

[54] J.L. Gauvain and C.H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc.* DARPA *Speech & Natural Language Workshop*, pp. 272-277, Pacific Grove, CA, February 1991.

[55] J.L. Gauvain and C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech & Audio Processing*, **2**(2), pp. 291-298, April 1994.

[56] E. Giachin, A.E. Rosenberg and C.H. Lee, "Word Juncture Modeling using Phonological Rules for HMM-based Continuous Speech Recognition,"*Computer Speech & Language*, **5**, pp. 155-168, *month* 1991.

[57] L. Gillick and R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *Proc.* DARPA *Speech & Natural Language Workshop*, pp. 170-172, Hidden Valley, PA, June 1990.

[58] L. Gillick, Y. Ito and J. Young, "A Probabilistic Approach to Confidence Measure Estimation and Evaluation", *Proc. IEEE ICASSP-97*, pp. 879-882, Munich, Germany, April 1997.

[59] J.R. Glass, T.J. Hazen and I. L. Hetherington, "Real-time Telephone-based Speech Recognition in the Jupiter Domain," *Proc. IEEE ICASSP-99*, **1**, pp. 61-64, Phoenix, AZ, March 1999.

[60] J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. IEEE ICASSP-92*, pp. 517-520, San Francisco, CA, March 1992.

[61] A. Goldschen and D. Loeh, "The Role of the DARPA Communicator Architecture as a Human Computer Interface for Distributed Simulations," *Proc. 1999 Simulation Interoperability Standards Organization (SISO) Spring Simulation Interoperability Workshop (SIW)*, Orlando, FL, March 14-19, 1999.

[62] P.S. Gopalakrishnan, L.R. Bahl and R.L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," *Proc. IEEE ICASSP-95*, **1**, pp. 572-575, Detroit, MI, May 1995.

[63] A.L. Gorin, G. Riccardi and J.H. Wright, "How may I help you?," *Speech Communication*, **23**(1-2), 113-127, October 1997.

[64] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland and S.J. Young, "Segment Generation and Clustering in the HTK Broadcast News Transcription System," *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 133-137, Landsdowne, VA February 1998.

[65] A.G. Hauptmann, M. Witbrock and M. Christel, "News-on-Demand-'An Application of Informedia Technology'," Digital Libraries Magazine, September 1995.

[66] C.T. Hemphill, J.J. Godfrey, and G.R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," *Proc.* DARPA *Speech & Natural Language Workshop*, Pittsburgh, PA, June 1990.

[67] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, **87**(4), pp. 1738-1752, 1990.

[68] M.M. Hochberg, S.J. Renals, A.J. Robinson and D. Kershaw, "Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system," in *Proc. ICSLP'94*, pp. 1499-1502, Yokohama, Japan, September 1994.

[69] M. Hwang and X. Huang, "Subphonetic Modeling with Markov States - Senone," *Proc. IEEE ICASSP-92*, San Francisco, CA, **1**, pp. 33-36, March 1992.

[70] M.Y. Hwang, X. Huang and F. Alleva, "Predicting Unseen Triphones with Senones," *Proc. IEEE ICASSP-93*, **II**, pp. 311-314, Minneapolis, MN, April 1993.

[71] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), pp. 532-556, April 1976.

[72] F. Jelinek, B. Merialdo, S. Roukos and M. Strauss, "A Dynamic Language Model for Speech Recognition," *Proc.* DARPA *Speech & Natural Language Workshop*, pp. 293-295, Pacific Grove, CA, February 1991.

[73] F. deJong, J.L. Gauvain, J. deb Hartog and K. Netter, "OLIVE: Speech Based Video Retrieval," *Proc. CBMI'99*, Toulouse, October 1999.

[74] P. Jourlin, S.E. Johnson, K. Spärck Jones and P.C. Woodland, "General Query Expansion Techniques for Spoken Document Retrieval," *Proc. SIGIR'99*, August 1999.

[75] C.M. Karat, C. Halverson, D. Horn and J. Karat, "Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition," *Proc. CHI'99*, 1999.

[76] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoustics, Speech & Signal Processing*, **ASSP-35**(3), pp. 400-401, March 1987.

[77] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. ESCA Eurospeech'99*, Budapest, Hungary, **6** 2725-2728, September 1999.

[78] D. Kershaw, A.J. Robinson and S.J. Renals, "The 1995 Abbot hybrid connectionist-HMM large-vocabulary recognition system," *Proc. ARPA Speech Recognition Workshop*, pp. 93-98, Harriman, NY, February 1996.

[79] R. Kneser and H. Ney, "Improved backing-off for n-gram language modeling," *Proc. IEEE ICASSP-95*, **1**, pp. 181-184, Detroit, MI, May 1995.

[80] F. Kubala, "Design of the 1994 CSR Benchmark Tests," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 41-46, Austin, TX, January 1995.

[81] F. Kubala, T. Anastasakos, H. Jin, J. Makhoul, L. Nguyen, R. Schwartz and N. Yuan, "Toward Automatic Recognition of Broadcast News," *Proc.* DARPA *Speech Recognition Workshop*, pp. 55-60, Harriman, NY, February 1996.

[82] L.F. Lamel and G. Adda, "On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition," *Proc. ICSLP'96*, **1**, pp. 6-9, Philadelphia, PA, October 1996.

[83] L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues and J.N. Temem, "User Evaluation of the MASK Kiosk," *Proc. ICSLP'98*, 2875-2878, Sydney, December 1998.

[84] L.F. Lamel and R. DeMori, "Speech Recognition of European Languages," *Proc. IEEE Automatic Speech Recognition Workshop*, pp. 51-54, Snowbird, Utah, December 1995.

[85] L.F. Lamel and J.L. Gauvain, "Continuous Speech Recognition at LIMSI," *Proc. ARPA Workshop on Continuous Speech Recognition*, pp. 59-64, Stanford, CA, September 1992.

[86] L.F. Lamel and J.L. Gauvain, "A Phone-based Approach to Non-Linguistic Speech Feature Identification," *Computer Speech & Language*, **9**(1), pp. 87-103, January 1995.

[87] L.F. Lamel, S. Rosset, S.K. Bennacef, H. Bonneau-Maynard, L. Devillers and J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information", *Proc. ESCA Eurospeech'95*, Madrid, Spain, **3**, pp. 1961-1964, Madrid, Spain, September 1995.

[88] L. Lamel, S. Rosset, J.L. Gauvain and S. Bennacef, "The LIMSI ARISE System," *Proc. IEEE IVTTA'98*, Torino, Italy, pp. 209-214, September 1998 (revised version to appear in *Speech Communication*.

[89] C.-H. Lee and Q. Huo, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition", these proceedings.

[90] K.-F. Lee, *Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system*, PhD Thesis, Carnegie Mellon University, 1988.

[91] L. Lee and R.C. Rose, "Speaker Normalisation Using Efficient Frequency Warping Procedures", *Proc. IEEE ICASSP-96*, **1**, pp. 353-356, Atlanta, GA, May 1996.

[92] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech & Language*, **9**, pp. 171-185, 1995.

[93] E. Levin and R. Pieraccini, "CHRONUS, The Next Generation," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, pp. 269-272, January 1995.

[94] A. Life, I. Salter, J.N. Temem, F. Bernard, S. Rosset, S. Bennacef and L. Lamel, "Data Collection for the MASK Kiosk: WOz vs Prototype System," *Proc. ICSLP'96*, pp. 1672–1675, Philadelphia, PA, October 1996.

[95] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, **22**(1), pp. 1-15, July 1997.

[96] D. Liu and F. Kubala, "Fast Speaker Change Detection for Broadcast News Transcription and Indexing.", *Proc. ESCA EuroSpeech'99*, **3**, pp. 1031-1034, Budapest, Hungary, September 1999.

[97] A. Ljolje, M.D. Riley, D.M. Hindle and F. Pereira, "The AT&T 60,000 Word Speech-To-Text System," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 162-165, Austin, TX, January 1995.

[98] Madcow, "Multi-site Data Collection for a Spoken Language Corpus," *Proc.* DARPA *Speech & Natural Language Workshop*, Harriman, NY, pp. 7-14, February 1992.

[99] B. Mak and E. Bocchieri, "Subspace distribution clustering for continuous observation density hidden Markov models," *Proc. Eurospeech'97*, pp. 107-110, Rhodes, Greece, September 1997.

[100] J.J. Mariani "Spoken Language Processing and Human-Machine Communication in the European Union Programmes," in G. Varile, ed., *Eurospeech'97 EU Speech Projects Day report*, Rhodes, Greece, September 1997.

[101] J.J. Mariani and L.F. Lamel, "An overview of EU programs related to conversational/interactive systems," *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 247-253, Landsdowne,VA, February 1998.

[102] M. Maybury (ed.), "News on Demand," Special section in the *Communications of the ACM* **43**(2), February 2000.

[103] D. Miller, R. Schwartz, R. Weischedel and R. Stone, "Named Entity Extraction from Broadcast News," *Proc.* DARPA *Broadcast News Workshop*, pp. 37-40, Herndon, VA, February 1999.

[104] W. Minker, "Evaluation Methodologies for Interactive Speech Systems," *Proc. LREC'98*, Granada, Spain, pp. 199–206, May 1998.

[105] W. Minker, "Stochastic versus rule-based understanding for information retrieval," *Speech Communication*, **25**(4), pp. 223-247, September 1998.

[106] M. Mohri, M. Riley, D. Hindle, A. Ljolie and F. Pereira, "Full Expansion of Context-Dependent Networks in Large Vocabulary Speech Recognition," *Proc. IEEE ICASSP-98*, pp. 665-668, Seattle, WA, May 1998.

[107] H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," *Proc. IEEE ICASSP-93*, **II**, pp. 319-322, Minneapolis, MN, April 1993.

[108] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, **ASSP-32**(2), pp. 263-271, April 1984.

[109] H. Ney, R. Haeb-Umbach, B.H. Tran and M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition," *Proc. IEEE ICASSP-92*, **I**, pp. 9-12, San Francisco, CA, March 1992.

[110] L. Nguyen and R. Schwartz, "Single-Tree Method for Grammar-Directed Search," *Proc. IEEE ICASSP-99*, **2**, pp. 613-616, Phoenix, AZ, March 1999.

[111] J.J. Odell, *The Use of Decision Trees with Context Sensitive Phoneme Modelling*, MPhil Thesis, Cambridge University Engineering Dept, 1992.

[112] J.J. Odell, V. Valtchev, P.C. Woodland and S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," *Proc. ARPA Human Language Technology Workshop*, pp. 405-410, Princeton, NJ, March 1994.

[113] E. den Os, L Boves, L. Lamel and P. Baggia, "Overview of the Arise Project," *Proc. ESCA Eurospeech'99*, **4**, 1527-1530, Budapest, Hungary, September 1999.

[114] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki and Z.P. Zeang, "Recent Advances in Japanese Broadcast News Transcription," *Proc. ESCA Eurospeech'99*, **2**, pp. 671-674, Budapest, Hungary, September 1999.

[115] S. Ortmanns, H. Ney, A. Eiden, "Language-model look-ahead for large vocabulary speech recognition," *Proc. ICSLP'96*, pp. 2095-2098, Philadelphia, PA, October 1996.

[116] B.T. Oshika, V.W. Zue, R.V. Weeks, H. Neu and J. Aurbach, "The Role of Phonological Rules in Speech Understanding Research," *IEEE Trans. Acoustics, Speech, Signal Processing*, **ASSP-23**, pp. 104-112, 1975.

[117] M. Ostendorf, A. Kannan, O. Kimball and J.R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proc. ARPA Workshop on Continuous Speech Recognition*, pp. 53-58, Stanford, CA, September 1992.

[118] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund and M.A. Pryzbocki, "1993 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Human Language Technology Workshop*, pp. 49-74, Princeton, NJ, March 1994.

[119] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A.F. Martin and M.A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 5-36, Austin, TX, January 1995.

[120] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, A.F. Martin and M.A. Przybocki, "1995 Hub-3 Multiple Microphone Corpus Benchmark Tests," *Proc. ARPA Speech Recognition Workshop*, pp. 27-46, Harriman, NY, February 1996.

[121] D.S. Pallett, J.G. Fiscus and M.A. Przybocki, "1996 Preliminary Broadcast News Benchmark Test," *Proc.* DARPA *Speech Recognition Workshop*, pp. 22-46, Chantilly, VA, February 1997.

[122] D.S. Pallett, J.G. Fiscus, A.F. Martin and M.A. Przybocki, "1997 Broadcast News Benchmark Test Results: English and Non-English," *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 5-11, Landsdowne, VA, February 1998.

[123] D.S. Pallett, J.G. Fiscus, J.S. Garofolo, A.F. Martin and M.A. Przybocki, "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures," *Proc.* DARPA *Broadcast News Workshop*, pp. 5-12, Herndon, VA, February 1999.

[124] D.B. Paul, "An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model," *Proc. IEEE ICASSP-92*, pp. 405-409, San Francisco, CA, March 1992.

[125] J. Peckham, "A New Generation of Spoken Dialog Systems: Results and Lessons from the SUNDIAL Project", *Proc. ESCA Eurospeech'93*, pp. 33-40, Berlin, Germany, September 1993.

[126] P. Price, "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc.* DARPA *Speech and Natural Language Workshop*, pp. 91-95, Hidden Valley, PA, June, 1990.

[127] M.K. Ravishankar, *Efficient Algorithms for Speech Recognition*, PhD Thesis, Carnegie Mellon University, 1996.

[128] F. Richardson, M. Ostendorf and J.R. Rohlicek, "Lattice-Based Search Strategies for Large Vocabulary Recognition," *Proc. IEEE ICASSP-95*, **1**, pp. 576-579, Detroit, MI, 1995.

[129] M.D. Riley and A. Ljojle, "Automatic Generation of Detailed Pronunciation Lexicons," *Automatic Speech and Speaker Recognition*, Kluwer Academic Pubs, Ch. 12, pp. 285-301, 1996.

[130] M.D. Riley, W. Byrne, M. Finke, S. Khudanpu, A. Ljojle, J. McDonough, H. Nock, M. Saraclar, C. Wooters and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Automatic Speech and Speaker Recognition, Speech Communication* **29**(2-4), pp. 209-224, November 1999.

[131] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Neaulieu and M. Gatford, "Okapi at TREC-3", *Proc. TREC-3*, Washington,DC, November 1994.

[132] A. Rogozan and P. Deléglise , "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Communication*, **26**(1-2), pp. 149-161, December 1998.

[133] R. Rosenfeld and X. Huang, "Improvements in Stochastic Language Modeling," *Proc.* DARPA *Workshop on Speech & Natural Language*, pp. 107-111, Harriman, NY, February 1992.

[134] R. Rosenfeld, *Adaptive Statistical Language Modeling*, PhD Thesis, Carnegie Mellon University, 1994. (also *Tech. rep. CMU-CS-94-138*)

[135] R. Rosenfeld, *Adaptive Statistical Language Modeling*, these proceedings.

[136] S. Rosset, S.K. Bennacef and L.F. Lamel, "Design Strategies for Spoken Language Dialog Systems," *Proc. ESCA Eurospeech'99*, **4**, pp. 1535-1538, Budapest, Hungary September 1999.

[137] A. Sankar, A. Stolke, T. Chung, L. Neumeyer, M. Weintraub, H. Franco and F. Beaufays, "Noise-Resistant Feature Extraction and Model Training for Robust Speech Recognition," *Proc. ARPA Speech Recognition Workshop*, pp. 117-122, Harriman, NY, February 1996.

[138] M. Schuster, "Memory-efficient LVCSR search using a one-pass stack decoder," *Computer Speech & Language*, **14**(1), pp. 47-77, January 2000.

[139] R. Schwartz, S. Austin, F. Kubala and J. Makhoul,"New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System," *Proc. IEEE ICASSP-92*, **I**, pp. 1-4, San Francisco, CA, March 1992.

[140] R. Schwartz, Y. Chow, S. Roucos, M. Krasner and J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition," *Proc. IEEE ICASSP-84*, **3**, pp. 35.6.1-35.6.4, San Diego, CA, March 1984.

[141] R. Schwartz, S. Miller, D. Stallard and J. Makhoul, "Language Understanding using Hidden Understanding Models," *Proc. ICSLP'96*, pp. **–**, Philadelphia, PA, October 1996.

[142] S. Sekine and R. Grishman, "NYU Language Modeling Experiments for the 1995 CSR Evaluation," *Proc. ARPA Speech Recognition Workshop*, pp. 123-128, Harriman, NY, February 1996.

[143] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *Proc. IEEE ICASSP-95*, pp. 697-700, Detroit, MI, May 1995.

[144] R. Schwartz, H. Jin, F. Kubala and S. Matsoukas, "Modeling Those F-Conditions – Or Not," *Proc.* DARPA *Speech Recognition Workshop*, pp. 115-118, Chantilly, VA, February 1997.

[145] K. Seymore and R. Rosenfeld, "Scalable backoff language models", *Proc. ICSLP'96*, **1**, pp. 232-235, Philadelphia, PA, October 1996.

[146] M. Siegler, U. Jain, B. Raj and R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *Proc.* DARPA *Speech Recognition Workshop*, pp. 97-99, Chantilly, VA, February 1997.

[147] M. Siu and H. Gish, "Evaluation of word confidence for speech recogni-

tion systems”, *Computer Speech & Language*, **13**(4), pp. 299-318, October 1999.

[148] A. Stolcke, “Entropy-based Pruning of Backoff Language Models”, *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 270-274, Landsdowne, VA, February 1998.

[149] G. Tajchman, E. Fosler and D. Jurafsky, “Building Multiple Pronunciation Models for Novel Words Using Exploratory Computational Phonology,” *Proc. ESCA Eurospeech’95*, **3**, pp. 2247-2250, Madrid, Spain, September 1995.

[150] S. Takahashi and S. Sagayama, “Four-level Tied Structure for Efficient Representation of Acoustic Modeling,” *Proc. IEEE ICASSP-95*, pp. 520-523, Detroit, MI, May 1995.

[151] L.F. Uebel and P.C. Woodland, “An Investigation into Vocal Tract Length Normalisation,” *Proc. ESCA Eurospeech’99*, pp. 2527-2530, Budapest, Hungary, September 1999.

[152] D.A. van Leeuwen, L.G. van den Berg and H.J.M. Steeneken, “Human Benchmarks for Speaker Independent Large Vocabulary Recognition Performance,” *Proc. ESCA Eurospeech’95*, pp. 1461-1464, Madrid, Spain, September 1995.

[153] T.K. Vintsyuk, “Speech discrimination by dynamic programming,” *Kibnernetika*, **4**, p. 81, *month* 1968.

[154] T.K. Vintsyuk, “Elements-wise recognition of continuous speech composed of words from a specified dictionary,” *Cybernetics*, **7**, pp. 133-143, March-April 1971.

[155] E. Voorhees and D. Harman, “Overview of the Eighth Text REtrieval Conference (TREC-8),” *Notebook of the 8th Text Retrieval Conference TREC-8*, pp. 1-15, Gaithersburg, MD, November 1999.

[156] W. Wahlster, “Verbmobil: Translation of Face-to-Face Dialogs,” *Proc. ESCA Eurospeech’93*, Berlin, Germany, **Plenary**, pp. 29-38, September 1993.

[157] F. Walls, H. Jin, S. Sista and R. Schwartz, “Probabilistic Models for Topic Detection and Tracking,” *Proc. IEEE ICASSP-99*, **1**, pp. 521-524, Phoenix, AZ, March 1999.

[158] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R.Roth and J. Yamron, “Dragon Systems’ 1997 Broadcast News Transcription System,” *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 60-65, Landsdowne, VA, February 1998.

[159] S. Wegmann, P. Zhan, L. Gillick, “Progress in Broadcast News Transcription at Dragon Systems,” *Proc. IEEE ICASSP’99*, pp. 33-36, Phoenix, AZ, March 1999.

[160] F. Wessel, K. Macherey and R. Schlüter, “Using word probabilities as confidence measures,” *Proc. IEEE ICASSP-98*, pp. 225-228, Seattle, WA, May 1998.

[161] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig and A. Stolcke, “Neural-Network based Measures of Confidence for Word Recognition,” *Proc. ICASSP-97*, pp. 887-890, Munich, Germnay, April 1997.

[162] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young, “The development of the 1994 HTK large vocabulary speech recognition system,” *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 104-109, Austin, TX, January 1995.

[163] P.C. Woodland, M.J.F. Gales, D. Pye and V. Valtchev, “The HTK large vocabulary recognition system for the 1995 ARPA H3 task,” *Proc. ARPA Speech Recognition Workshop*, pp. 99-104, Harriman, NY, February 1996.

[164] J.P. Yamron, I. Carp, L. Gillick, S. Lowe and P. van Mulbregt, “A Hidden Markov Approach tp Text Segmentation and Event Tracking, *Proc. IEEE ICASSP-98*, **1**, pp. 333-336, Seattle, WA, May 1998.

[165] S.J. Young, “The General Use of Tying in Phoneme-Based HMM Speech Recognisers,” *Proc. IEEE ICASSP-92*, San Francisco, CA, pp. 569-572, March 1992.

[166] S.J. Young, “A Review of Large-Vocabulary Continuous Speech Recognition,” *IEEE Signal Processing Magazine*, **13**(5), pp. 45-57, September 1996.

[167] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, H.J.M. Steeneken, A.J. Robinson and P.C. Woodland, “Multilingual large vocabulary speech recognition: the European SQALE project,” *Computer Speech & Language*, **11**(1):73-89, January 1997.

[168] S.J. Young and L. Chase, “Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes,” *Computer Speech & Language*, **12**(4), pp. 263-279, October 1998.

[169] S.J. Young, J.J. Odell and P.C. Woodland, “Tree-Based State Tying for High Accuracy Acoustic Modeling,” *Proc. ARPA Human Language Technology Workshop*, pp. 307-312, Princeton, NJ, March 1994.

[170] S.J. Young and P.C. Woodland, “The Use of State Tying in Continuous Speech Recognition,” *Proc. ESCA Eurospeech’93*, **3**, pp. 2203-2206, Berlin, Germany, September 1993.

[171] G. Zavaliagkos, R. Schwartz and J. McDonough, “Maximum *a Posteriori* Adaptation for Large Scale HMM Recognizers,” *Proc. IEEE ICASSP-95*, pp. 725-728, Detroit, MI, May 1995.

[172] V. Zue. J. Glass, M. Phillips and S. Seneff, “The MIT SUMMIT Speech Recognition System: A Progress Report”, *Proc.* DARPA *Speech & Natural Language Workshop*, pp. 179-189, Philadelphia, PA, February 1989.