chapter 1

LARGE VOCABULARY SPEECH RECOGNITION BASED ON STATISTICAL METHODS

Introduction

Speech recognition is concerned with converting the speech waveform, an acoustic signal, into a sequence of words. Today's most performant approaches are based on a statistical modelization of the speech signal. The chapter provides an overview of the main topics addressed in large vocabulary speech recognition, that is acoustic-phonetic modeling, lexical representation, language modeling, decoding and model adaptation. Only a few years ago speech recognition was primarily associated with a limited number of applications: small vocabulary isolated word recognition 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.

This chapter focuses on methods used in state-of-the-art speaker-independent, large vocabulary continuous speech recognition. Some of the primary application areas for LVCSR technology are dictation, spoken language dialog, and transcription systems for information archival and retrieval. After providing an overview of LVCSR, detailed discussions of statistical methods for each of the system components are given. Some outstanding issues and directions of future research are discussed.

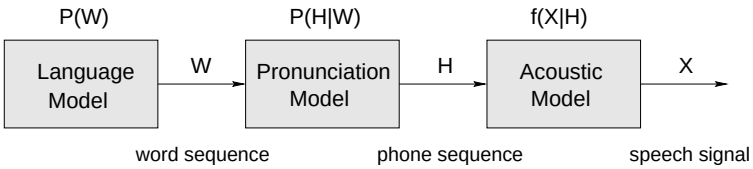

Figure 1.1 LVCSR speech generation model: The word sequence W produced by the language model is successively transformed by the pronunciation model ($P(H|W)$) and the acoustic model ($f(X|H,W)$), resulting in the speech signal X.

LVCSR Overview

From a statistical point of view, speech is assumed to be generated by a language model which provides estimates of $\Pr(W)$ for all word strings $W = (w_1, w_2, ...)$ and an acoustic model encoding the message W in the signal X, represented by a probability density function $f(X|W)$. The goal of speech recognition is generally defined as finding the most likely word sequence given the observed acoustic signal. The speech decoding problem thus consists of maximizing the probability of W given the speech signal X, or equivalently, maximizing the product $\Pr(W)f(X|W)$.

LVCSR systems use acoustic units corresponding to phone or phone-in-context units,[1] where each word is described by one or more phone transcriptions. Assuming that the speech signal X depends only on the underlying phone sequence $H = (h_1, h_2, ...)$, then the expression $f(X|W)$ can be rewritten as $\sum_H \Pr(H|W)f(X|H)$ where the summation is taken over the set pronunciations corresponding to the word sequence W. (In practice the set is reasonably small as the average number of pronunciation variants per word is less than 2.) The underlying speech generation model is illustrated in Figure 1.1. The word sequence produced by the language model is successively transformed by two transducers, the pronunciation model and the acoustic model, to yield the audio signal.

This formulation of the LVCSR problem leads to the following 4 main considerations:

- the language modeling problem, i.e. computing the a priori probability $Pr(W)$. This is usually estimated from n-gram frequencies in text corpora and transcriptions of speech data,

- the pronunciation modeling problem, i.e. the computation of $Pr(H|W)$, which relies on a pronunciation dictionary and may include estimates of the word pronunciation probabilities,

[1] In this chapter the term phone is used instead of phoneme (refering to the elementary and distinctive sounds in the language) or phonetic (the observed realization of the elementary sounds). Contextual phone units (phone-in-context) implicitly model what can be consider allophones, i.e. contextual variants of the underlying phoneme.

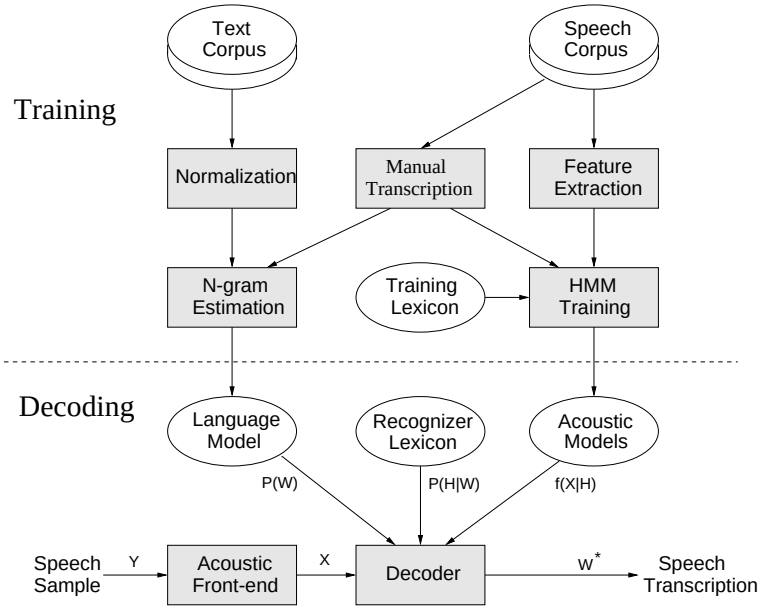

Figure 1.2 System diagram of a generic speech recognizer based on statistical models, including training and decoding processes and the main knowledge sources.

- the acoustic modeling problem, i.e. determining the structure of the probability density function $f(X|H)$ and estimating its statistical parameters from speech samples. The most predominant approach uses continuous density hidden Markov models (HMM) to represent context-dependent phones.

- the search problem, i.e. determining the best word hypothesis for the speech data given the models. This is a big challenge for LVCSR due to the large vocabulary and language model sizes.

The principles on which most state-of-the-art LVCSR systems are based have been known for many years now, and include the application of the communication theory to speech recognition[11, 78, 79], the use of a spectral representation of the speech signal[38, 39], the use of dynamic programming for decoding [166, 167], and the use of context-dependent acoustic models [28, 100, 153]. Despite the fact that some of these techniques were proposed well over a decade ago, considerable progress has been made in recent years in part due to the availability of large speech and text corpora, and improved processing power which have allowed more complex models and algorithms to be implemented. Compared with the state-of-the-art technology a decade ago, advances in acoustic modeling and model adaptation have enabled reasonable performance to be obtained on various data types and acoustic conditions.

The main components of a generic speech recognition system are shown in Figure 1.2. The elements shown are the main knowledge sources (speech and

textual training materials and the pronunciation lexicon), the feature analysis (or parameterization), the acoustic and language models which are estimated in a training phase, and the decoder. The remainder of this chapter is devoted to discussing these main constituents and knowlegde sources.

Language modeling

Language models (LMs) capture regularities in spoken language and are used in speech recognition to estimate the probability of word sequences. Grammatical constraints can be described using a context-free grammars (for small to medium size vocabulary tasks these are usually manually elaborated) or can be modeled stochastically, as is common for LVCSR. The most popular statistical methods are n-gram models, which attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of n words. The assumption is made that the probability of a given word string $W = (w_1, w_2, ..., w_k)$ can be approximated by the following forward sequential decomposition

$$P(W) = \prod_{i=1}^{k} \Pr(w_i | w_{i-n+1}, ..., w_{i-2}, w_{i-1})$$

therefore reducing the word history to the preceding $n-1$ words. It should be noted that other decompositions of $P(W)$ can also be appropriate, for example, a backward decomposition will lead to a backward n-gram model.

A prerequisite for estimating n-gram language models is the availability of appropriately processed text corpora. As can be seen in Figure 1.2, language models are usually estimated from manual transcriptions of speech corpora and from normalized text corpora. To ensure accurate models, the texts should be as representative as possible of the expected audio input to be transcribed. Text preparation entails locating appropriate sources of text data and audio transcriptions, and processing them in a homogeneous manner. The recognizer vocabulary (also called word list) is selected usually with the goal of maximizing lexical coverage, and the n-gram probabilities are estimated using appropriate smoothing techniques. Language models are generally optimized and compared by measuring the likelihood of a set of left out data, referred to as LM development texts or development data. The relevance of a language model in terms of test set perplexity is defined as:

$$\mathrm{Px}(T|M) = P(T|M)^{-\frac{1}{L}} \simeq (\prod_{i=1}^{L} P(w_i | w_{i-2}, w_{i-1}))^{-\frac{1}{L}}$$

for a given text $T = (w_1, ..., w_L)$ and a language model M. $P(T|M)$ denotes the language model estimate of the text probability. The test set perplexity depends on both the language being modeled and the model, i.e., it gives a combined estimate of how good model is and how complex the language

is [79]. If the text set is representative of the model, the perplexity can be seen as a measure of the average branching factor, i.e. the vocabulary size of a memoryless uniform language model with same entropy as the language model under consideration.

Text preparation

Although ideal language model training data would consist of large corpora of transcribed audio data for a particular task, in practice such data are difficult to obtain. Therefore a variety of other more or less closely related text materials are usually used for language model training.

Given a large text corpus 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 [29]. 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 back-off 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. Much of speech recognition research for American English has been supported by ARPA and has been based on text materials which were processed to remove upper/lower case distinction and compounds. Thus, for instance, no lexical distinction is made between *Rich, rich* or *Brown, brown*. In the French *Le Monde* corpus, capitalization of proper names is distinctive with different lexical entries for *Pierre (proper name), pierre (stone)* or *Roman (proper name), roman (novel)*.

The main conditioning steps are text mark-up and conversion. Text mark-up consists of tagging the texts (article, paragraph and sentence markers) and garbage bracketing (which includes not only corrupted text materials, but all text material unsuitable for sentence-based language modeling, such as tables and lists). Numerical expressions and dates are typically expanded to approximate the spoken form and to reduce the lexical variety ($150 $\rightarrow$ one hundred and fifty dollars, 1991 $\rightarrow$ nineteen ninety one or one thousand nine hundred and ninety one). Further semi-automatic processing is necessary to correct frequent errors inherent in the texts (such as obvious mispellings *milllion*, *officals*) or arising from processing with the distributed text processing tools. Some normalizations can be considered as "decompounding" rules in they modify the word boundaries and the total number of words. These concern the processing of ambiguous punctuation markers (such as hyphen and apostrophe), the processing of digit strings, and treatment of abbreviations and acronyms (ABCD $\rightarrow$ A. B. C. D.). While such expansions increase lexical coverage in some languages such as English and French, in German, the standard written form of the dates are formed by aggluinating the component words. For example, 1991 is written as *neunzehnhunderteinundneunzig*. De-

HUNDRED <nb>	$\Longrightarrow$	HUNDRED AND <nb> (0.50)
ONE EIGHTH	$\Longrightarrow$	AN EIGHTH (0.50)
CORPORATION	$\Longrightarrow$	CORP. (0.29)
INCORPORATED	$\Longrightarrow$	INC. (0.22)
ONE HUNDRED	$\Longrightarrow$	A HUNDRED (0.19)
MILLION DOLLARS	$\Longrightarrow$	MILLION (0.15)
BILLION DOLLARS	$\Longrightarrow$	BILLION (0.15)

Figure 1.3 Some example transformation rules applied during text normalization with associated probabilities.

compounding rules can be used to transform this date into the word sequence *neunzehn hundert ein und neunzig*, thereby reducing lexical variety. Depending upon the target application, the recognizer hypotheses may need to be mapped to a more appropriate written form. Other normalizations (such as sentence initial capitalization and case distinction) keep the total number of words unchanged, but reduce graphemic variability. In general the choice is a compromise between producing an output close to correct standard written form of the language and lexical coverage, with the final choice of normalization being largely application-driven.

Better models of spoken language can be obtained by transforming text data to be closer to an oral form. In the case of read speech corpora, such as can be used for dictation tasks, the transformation rules and corresponding probabilities can be automatically derived by aligning transcriptions with the printed text form. Some example transformations are shown in Figure 1.3 along with the rule probabilites. For example, the word HUNDRED followed by a number can be replaced by *hundred and* 50% of the time; 50% of the occurences of *one eighth* are replaced by *an eighth*, and 15% of the sequence *million dollars* are replaced with simply the word *million* [58].

Vocabulary selection

In practice, the selection of words is done so as to minimize the system's out-of-vocabulary (OOV) rate by including the words which are expected to be most frequent in the input. These words must also be sufficiently frequent in the available text corpora in order to be able to train a language model. This condition is usually met by choosing the N most frequent words in the training data. This criterion does not, however, guarantee the usefulness of the vocabulary, since no consideration of the expected input is made. It is therefore common practice to use the LM development data to select a word list adapted to the expected test conditions.

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 [24, 58]. On average, each OOV word causes more than a single error, with rates of 1.6 to 2.0 additional errors reported [131]. 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 [58], so in contradiction to the widely held belief, larger vocabulary does not imply higher word error rates if a proper language model is used.

N-gram Estimation

Using the maximum likelihood (ML) criterion, the n-gram probabilities can be estimated from the frequencies of the word sequences of length n in the training corpus (texts or speech transcriptions). For example, the ML estimate of the trigram probability is given by:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}$$

where $C(\cdot)$ is the number of times the n-gram appears in the training data. However, obtaining reliable probability estimates requires multiple observations of the n-gram. This requirement is usually met by only modeling word sequences that occur at least a minimal count number of times.

Estimating the n-gram probabilities requires sufficient quantities of representative text materials. For large vocabulary sizes, many of the possible n-grams will not occur in even a very large training corpus. Therefore estimates that rely on counts of observed n-grams will be unable to predict many possible word sequences, and due to the sparseness of the data, the maximum likelihood estimates are inadequate and need to be smoothed. Different approaches have been investigated to smooth the estimates of the probabilities of rare n-grams [26, 89]. The most commonly used approach is to apply a back-off mechanism [85] which relies on a lower order n-gram when there is insufficient training data. The back-off also provides a means of modeling unobserved word sequences. For example, if there is not enough data to obtain a robust estimate from the n-gram counts, a fraction of the probability mass is taken from the observed n-grams by discounting the ML estimates [67, 89, 175]. The probabilities of the rare n-grams are then estimated from the $(n-1)$-gram probabilities in a recursive manner shown here for a trigram model:

$$\hat{P}(w_i|w_{i-2}, w_{i-1}) = \hat{P}(w_i|w_{i-1})\mathrm{B}(w_{i-2}, w_{i-1}),$$

where $\mathrm{B}(w_{i-2}, w_{i-1})$ is a back-off coefficient needed to ensure that the probability sum for a given context is always equal to 1. Computing the bigram estimate $\hat{P}(w_i|w_{i-1})$ follows the same principle. Backing-off offers an additional advantage in that the language model size can be arbitrarily reduced

by increasing the cutoff frequencies below which the n-grams are not included in the model. This property can also be used to reduce the amount of computation required during decoding. While 2-gram and 3-gram LMs are the most widely used, small improvements have been reported with the use of longer span 4-grams [12, 176] and 5-grams [108] or word class-based 5-grams [150].

It is often the case that different types of LM training material are available in differing quantities, and in different formats. A first step in combining sources means carrying out common normalizations. There are two commonly used approaches to estimate language models on different data sources: combining the models or merging the data. Model interpolation is an easy way to combine training material from different sources. A language model is trained for each source and the resulting models are interpolated. 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 the development data set. 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 back-off coefficients.

While trigram LMs are the most widely used, higher order (n>3) and word class-based (counts are based on sets of words rather than individual lexical items) n-grams, and adapted LMs are recent research areas aimed at improving LM accuracy. Class or category-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.

Given some training data and a mapping which assigns each word a unique category $\mathcal{C}(w)$, the training text can tagged and the n-gram probabilities $\Pr(w_i|\mathcal{C}(w_{i-n+1}), ..., \mathcal{C}(w_{i-1}))$, which are often approximated by $\Pr(w_i|\mathcal{C}(w_i)) \Pr(\mathcal{C}(w_i)|\mathcal{C}(w_{i-n+1}), ..., \mathcal{C}(w_{i-1}))$, can be estimated from the relative frequencies as is done for a regular word n-gram LM. The class assignement is often obtained by minimizing the perplexity of a bigram category model for a given number of word categories [88, 113]. In order to obtain a lower perplexity than that obtained with a regular n-gram model, it is wise to interpolate the category LM with the n-gram LM. The resulting trigram probability estimates are:

$$P^*(w_i|w_{i-2}, w_{i-1}) = \alpha \hat{P}(w_i|w_{i-2}, w_{i-1}) + (1-\alpha)\hat{P}(w_i|\mathcal{C}(w_{i-2}), \mathcal{C}(w_{i-1}))$$

Other statistical language models have been investigated which essentially map the word history $(w_1, ..., w_{i-1})$ onto equivalence classes rather than $(n-1)$-grams. These modeling techniques such decision tree models, maximum entropy models, and linguistically motivated models such probabilistic context-

free and link grammars, have been used with moderate success leading to small gains over the much simpler n-gram models [148].

LM Adaptation

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 dynamic, dependent for example, on the dialog state. Language model adaptation is of interest for improving the model accuracy and for keeping the models up-to-date. This is of particular importance for tasks such as broadcast news transcription in which news topics appear suddenly, and remain popular for variable lengths of time. Various approaches have been taken to adapt the language model based on the observed text so far, including the use of a *cache model* [80, 147], a *trigger model* [146], or *topic coherence modeling* [155]. 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.

Pronunciation modeling

The pronunciation dictionary is the link between the acoustic-level representation and the word sequence output by the speech recognizer. Designing a pronunciation dictionary has two main aspects: definition and selection of the vocabulary items and representation of each pronunciation entry using the basic acoustic units of the recognizer. Recognition performance is obviously related to lexical coverage, and the accuracy of the acoustic models is linked to the consistency of the pronunciations associated with each lexical entry. As discussed above, recognition vocabulary is usually selected to maximize lexical coverage for a given size lexicon. Judicious selection of the recognition vocabulary is important since on average, each out-of-vocabulary word causes more than one error (usually between 1.5 and 2 errors),

Associated with each lexical entry are one or more pronunciations, described using the chosen elementary units (usually phonemes or phone-like units). This set of unit is evidently language dependent. For example, some commonly used phone set sizes are about 45 for English, 50 for German and Italian, 35 for French and Mandarin (to which tones may be added), and 25 for Spanish. In generating pronunciation baseforms, most lexicons

Phone	*Example*	*Phone*	*Example*
Vowels		Fricatives	
i	b<u>ee</u>t	s	<u>s</u>ue
ɪ	b<u>i</u>t	z	<u>z</u>oo
e	b<u>ai</u>t	ʃ	<u>sh</u>oe
ɛ	b<u>E</u>t	ʒ	mea<u>s</u>ure
æ	b<u>a</u>t	f	<u>f</u>an
ʌ	b<u>u</u>t	v	<u>v</u>an
ɑ	b<u>o</u>tt	θ	<u>th</u>in
ɔ	b<u>ough</u>t	ð	<u>th</u>at
o	b<u>oa</u>t	Plosives	
u	b<u>oo</u>t	b	<u>b</u>et
U	b<u>oo</u>k	d	<u>d</u>ebt
ɝ	b<u>ir</u>d	ɡ	<u>g</u>et
Diphthongs		p	<u>p</u>et
ɑʲ	b<u>i</u>te	t	<u>t</u>at
ɔʲ	b<u>oy</u>	k	<u>c</u>at
ɑʷ	b<u>ou</u>t	Affricates	
Reduced Vowels		ʧ	<u>ch</u>eap
ə	<u>a</u>bout	ʤ	<u>j</u>eep
ɨ	dat<u>e</u>d	Nasals	
ɚ	butt<u>er</u>	m	<u>m</u>et
Semivowels		n	<u>n</u>et
l	<u>l</u>ed	ŋ	thi<u>ng</u>
r	<u>r</u>ed	Syllabics	
w	<u>w</u>ed	m̩	bott <u>om</u>
y	<u>y</u>et	n̩	butt<u>on</u>
h	<u>h</u>at	l̩	bott<u>le</u>

Figure 1.4 Set of phone symbols for English with illustrative words.

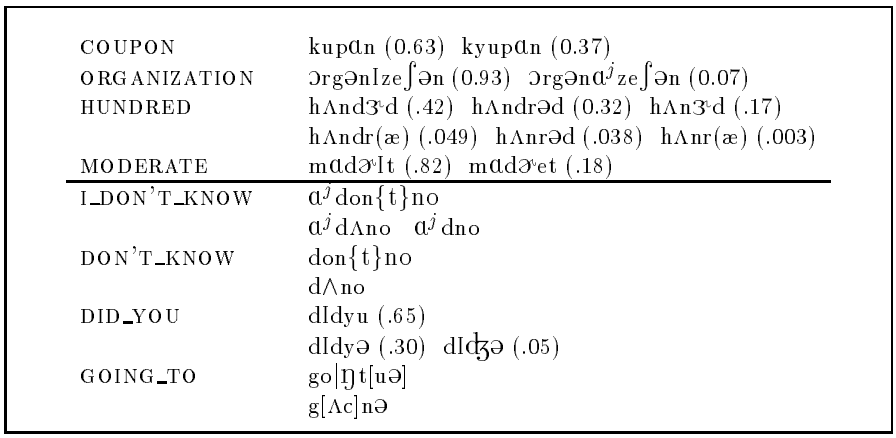

Figure 1.5 Some example lexical entries and their pronunciations along with estimate probabilities. For the compound words, the original concatenated pronunciation is given in the 1st line and the reduced forms are given in the 2nd line. Phones in ( ) specify the context for the pronunciation.

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. While pronunciation lexicons are usually (at least partially) created manually, several approaches to automatically learn and generate word pronunciations have been investigated[25, 30, 45, 142, 162]. To the best of our knowledge such approaches, while promising, have to date, given only small performance improvements even when trained with manual transcriptions [143].

Pronunciation variants which are not allophonic differences can be observed for a variety of words. Alternative pronunciations are evidently needed for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as *excuse, record, moderate*. Some frequent affixes such as *anti-, bi-, multi-, -ization* can be pronounced with a diphthong (/ɑ^i/) or a short vowel (/I/ or /ə/). Words of foreign origin, particularly proper names, may have different pronunciations depending upon the speaker's familiarity with the original language. It is also common for multisyllablic words to be pronounced with fewer syllables than in the full form, such as *company, interest, conference*. If acoustic model training is carried out without allowing for appropriate pronunciation variants, there will necessarily be a misalignment of one or more phones, adding noise to the phone

models. An optimal alignment with a pronunciation dictionary including all required variants results in more accurate acoustic phone models. Experience has shown that careful lexical design can improve speech recognition system performance [92].

In speech from fast speakers or speakers with relaxed speaking styles it is common to observe poorly articulated (or skipped) unstressed syllables, 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 in the lexicon can allow schwa-deletion or syllabic consonants in unstressed syllables. Compound words have also been used as a way to represent reduced forms for common word sequences such as "don't know" (dunno), "did you" (dija) and "going to" (gonna). Some example compound words are shown in Figure 1.5 along with estimates of the pronunciation probabilites for the different variants. Alternatively, such fluent speech effects can be modeled using phonological rules [128]. The principle behind the phonological rules is to modify the allowable phone sequences to take into account such variations. These rules are 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 has been used to handle liaisons, mute-e, and final consonant cluster reduction for French [57].

As speech recognition research has moved from read speech to found audio data, the phone set has been expanded to include non-speech events. These can correspond to noises produced by the speaker (breath noise, coughing, sneezing, laughter, etc.) or can correspond to external sources (music, motor, tapping etc).

Acoustic Modeling

Acoustic parameterization is concerned with the choice and optimization of acoustic features in order to reduce model complexity while trying to maintain the linguistic information relevant for speech recognition. Acoustic modeling must take into account different sources of variability present in the speech signal: those arising from the linguistic context and those associated with the non-linguistic context such as the speaker (e.g., coughing, throat clearing, breath noise) and the acoustic environment (e.g., background noise, music) and recording channel (e.g., direct microphone, telephone). Most state-of-the-art systems make use of hidden Markov models for acoustic modeling [139, 180], which consists of modeling the probability density function of a sequence of acoustic feature vectors. Other approaches include segment based models [64, 129, 186] and neural networks [5, 74] 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.

In this section common parameterizations are described, followed by a discussion of acoustic model estimation and adaptation.

Acoustic front-end

The first step of the acoustic feature analysis is digitization, where the continuous speech signal is converted into discrete samples. The most commonly used sampling rates are 16kHz and 10kHz for direct microphone input, and 8kHz for telephone signals. The next step is feature extraction (also called parameterization or front-end analysis), which has the goal of representing the audio signal in a more compact manner by trying to remove redundancy and reduce variability, while keeping the important linguistic information [75]. Most recognition systems use short-time cepstral features based either on a Fourier transform or a linear prediction model. Cepstral parameters are popular because they are a compact representation, and are less correlated than direct spectral components. This simplifies estimation of the acoustic model parameters by reducing the need for modeling the feature dependency. An inherent assumption is that although the speech signal is continually changing, due to physical constraints on the rate at which the articulators can move, the signal can be considered quasi-stationary for short periods (on the order of 10ms to 20ms).

The two most popular sets of features are cepstrum coefficients obtained with a Mel Frequency Cepstral (MFC) analysis [32] or with a Perceptual Linear Prediction (PLP) analysis [73]. In both cases a Mel scale short term power spectrum is estimated on a fixed window (usually in the range of 20 to 30ms). In order to avoid spurious high frequency components in the spectrum due to discontinuities caused by windowing the signal, it is common to use a tapered window such as a Hamming window. The window is then shifted (usually a third or a half the window size), and the next feature vector computed. The most commonly used offset is 10ms. The acoustic parameterization converts the speech signal Y into a sequence of feature vectors X, each vector representing a 10 ms interval:

$$X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T).$$

The Mel scale approximates the frequency resolution of the human auditory system, being linear in the low frequency range (below 1000 Hz) and logarithmic above 1000 Hz. The cepstral parameters are obtained by taking an inverse transform of the log of the filterbank parameters. In the case of the MFC coefficients, a cosine transform is applied to the log power spectrum, whereas a root-Linear Predictive Coding (LPC) analysis is used to obtain the PLP cepstrum coefficients. Both set of features have been used with success for LVCSR, but PLP analysis has been found for some systems to be more robust in presence of background noise [87, 177]. Finding the optimal tuning, which may be dependent on the language or the channel conditions, can

result in slight performance improvements. The set of cepstral coefficients associated with a windowed portion of the signal is referred to as a frame or a parameter**feature** vector. Cepstral mean removal (subtraction of the mean from all input frames) [46] is commonly used to reduce the dependency on the acoustic recording conditions. Computing the cepstral mean requires that all of the signal is available prior to processing, which is not the case for certain applications where processing needs to be synchronous with recording. In this case, a modified form of cepstral subtraction can be carried out where a running mean is computed from the N last frames (N is often on the order of 100, corresponding to 1s of speech). It is also common to normalize the variance, so that each resulting cepstral coefficient therefore has a zero mean and unity variance. In order to capture the dynamic nature of the speech signal, it is common to augment the feature vector with "delta" parameters. The delta parameters are computed by taking the first and second differences of the parameters in successive frames. As a result a typical feature vector $\mathbf{x}_t$ will include 12 cepstrum coefficients and the normalized log-energy, along with the first and second order derivatives, i.e. a total of 39 components.

***Acoustico-phonetic** Allophone modeling*

**there are many repeats in this section **

Hidden Markov models are widely used to model the sequences of acoustic feature vectors [139]. These models are popular as they are performant and their parameters can be efficiently estimated using well established techniques. They are used to model the production of speech feature vectors in two steps. First a Markov chain is used to generate a sequence of states, and second speech vectors are drawn using a probability density function (PDF) associated to each state. The Markov chain is described by the number of states and the transitions probabilities between states. Most recognition systems use acoustic units corresponding to phone or phone-in-context units. However it is 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 [13]. 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 back-off mechanisms to model infrequent contexts. The most widely used elementary acoustic units in LVCSR systems are phone-based, where each phone[2] is represented by a Markov chain with a small number of

[2]Phones usually correspond to phonemes, but may also correspond to allophones such as flaps or glottal stop.

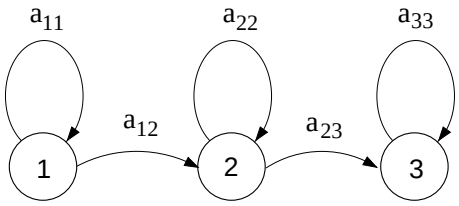

Figure 1.6 A simple 3-state left-to-right HMM topology commonly used for allophone modeling in LVCSR.

states. While different topologies have been proposed, all make use of left-to-right state sequences in order to capture the spectral change across time. The most commonly used configurations have between 3 and 5 emitting states per model, where the number of states imposes a minimal time duration for the unit. Some configurations allow certain states to be skipped, so as to reduce the required minimal duration. The probability of an observation (i.e. a speech vector) is assumed to be dependent only on the state, which is known as a 1st order Markov assumption.

Strictly speaking, given an N-state HMM with parameter vector λ, the HMM stochastic process is described by the following joint probability density function $f(X, S|\lambda)$ of the observed signal $X = (\mathbf{x}_1, ..., \mathbf{x}_T)$ and the unobserved state sequence $S = (s_0, ..., s_T)$,

$$f(X, S|\lambda) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} f(\mathbf{x}_t|s_t)$$

where π_i is the initial probability of state i, a_{ij} is the transition probability from state i to state j, and $f(\cdot|s)$ is the emitting PDF associated with each state s. Figure 1.6 shows the transition structure of a 3-state left-to-right commonly used for allophone modeling in LVCSR. Such model generate at 3 speech frames per allophone, making the minimal duration of a phone segment equal to 30ms for frame rate of 100Hz.

The most commonly used state output PDF for speaker-independent LVCSR systems is a mixture of Gaussians with about 16 to 32 components, thus

$$f(\mathbf{x}_t|s) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}_t|\mathbf{m}_{sk}, \Sigma_{sk})$$

where $\mathbf{m}_{sk}$, Σ_{sk} and ω_k denotes respectively the mean vector, the covariance matrix and the mixture weight of the k-th Gaussian component of the state s. To reduce the number of parameters and the inherent estimation problem linked to full covariance matrices, the covariance matrices are usually assumed diagonal but some recent developpment have demonstrated that non diagonal covariance matrices can be used while keeping the estimation problem managable [**?**, **?**].

Phone based models offer the advantage that recognition lexicons can be described using the elementary units of the given language, and thus benefit

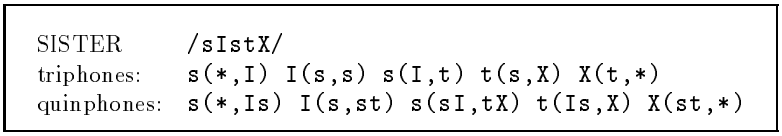

Figure 1.7 Exemples of allophonic transcriptions for intra-word triphones and quinphones. Each contextual unit is defined by the central phone followed by its phonemic context shown in parentheses (left-context, right-context).

from many linguistic studies. It is of course possible to perform speech recognition without using a phonemic lexicon, either by use of "word models" (as was the more commonly used approach 10 years ago) or a different mapping such as the fenones [13]. Compared with larger units (such as words, syllables, demisyllables), small subword units reduce the number of parameters, enable cross word modeling, facilitate porting to new vocabularies and most importantly, can be associated with back-off mechanisms to model rare contexts. Fenones offer the additional advantage of automatic training, but lack the ability to include *a priori* linguistic models.

A given HMM can represent a phone without consideration of its neighbors (context-independent or monophone model) or a phone in a particular context (context-dependent model).

***some duplicate text around**

Various types of contexts have been investigated from a single phone context (right- or left-context), left and right-context (triphone), generalized triphones [100], position-dependent triphones (cross-word and within word triphones), function word triphones, and quinphones [176]. The context may or may not include the position of the phone within the word (word-position dependent), and word-internal and cross-word contexts may be merged or considered separated models. The use of cross-word contexts complicates decoding, as discussed below. Different approaches are used to select the contextual units based on frequency or using clustering techniques, or decision trees, and different context types have been investigated: single-phone contexts, triphones, generalized triphones, quadphones and quinphones, with and without position dependency (within or cross word).

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.

Using contextual phone models can be seen as replacing the phonemic transcriptions from the pronunciation dictionary by allophonic transcriptions. Fig-

ure 1.7 gives the triphone and quinphone transcriptions for the word SISTER assuming only word internal units are used, i.e. the allophonic transcription are independent of the word context. When using cross-word triphones the models used for the first and last phone of each word (this is extended to the first and last two phones in the case of quinphones) will depend on the word context making the decoding problem significantly more complex.

The model states are often clustered so as to reduce the model size, resulting in what are referred to as "tied-state" models. **to be developped (see IEEE paper and French paper)***

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* [76], *genones* [35], *PELs* [16], *tied-states* [184]). Numerous ways of tying HMM parameters have been investigated [163, 179] 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 [77, 183]. The set of questions typically concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones [123].

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.

HMM Parameter Estimation

Acoustic model training consists of estimating the parameters of each HMM. For continuous density Gaussian mixture HMMs, this requires estimating the means and covariance matrices, the mixture weights and the transition probabilities. The most popular approaches make use of the Maximum Likelihood (ML) criterion, ensuring the best match between the model and the training data (assuming that the size of the training data is sufficient to provide robust estimates).

Estimation of the model parameters is usually done with the Expectation-Maximization (EM) algorithm [33] which is an iterative procedure starting with an initial set of model parameters. The model states are then aligned to the training data sequences and the parameters are reestimated based on this new alignment using the Baum-Welch reestimation formulas [17, 105, 83]. This algorithm guarantees that the likelihood of the training data given the models increases at each iteration. In the alignment step a given speech frame can be assigned to multiple states (with probabilities summing to 1) using the forward-backward algorithm or to a single state (with probability 1) using the Viterbi algorithm. This second approach yield to slightly lower likelihood but in practice there is very little difference in accuracy especially when large amounts of data are available. It is important to note that the EM algorithm does not guaranty finding the true ML parameter values, and even when the true ML estimates are obtained they may not be the best ones for speech recognition. Therefore, some implementation details such as a proper initialization procedure and the use of constraints on the parameter values can be quite important.

Since the goal of training is to find the best model to account of the observed data, the performance of the recognizer is critically dependent upon the representativity of the training data. Some methods to reduce this dependency are discussed in the next subsection. Speaker-independence is obtained by estimating the parameters of the acoustic models on large speech corpora containing data from a large speaker population. There are substantial differences in speech from male and female talkers arising from anatomical differences (on average females have a shorter vocal tract length resulting in higher formant frequencies, as well as a higher fundamental frequency) and social ones (female voice is often "breathier" caused by incomplete closure of the vocal folds). It is thus common practice to use separate models for male and female speech in order to improve recognition performance. The sex-dependent models are often obtained from speaker-independent seed models using Maximum *A Posteriori* estimators [60], or may be trained on the independent data subsets if sufficient training data are available. This choice evidently requires automatic identification of the gender.

MMIE

HMM Adaptation

In this section we discuss techniques that have been used with continuous density HMMs, although similar techniques have been developed for discrete and semi-continuous HMMs.

The performances of speech recognizers drop substantially when there is a mismatch between training and testing conditions. Several approaches can be used to minimize the effects of such a mismatch, so as to obtain a recognition accuracy as close as possible to that obtained under matched conditions. Acoustic model adaptation can be used to compensate mismatches between the training and testing conditions, such as due to differences in acoustic environment, to microphones and transmission channels, or to particular speaker characteristics. The techniques are commonly referred to as noise compensation, channel adaptation, and speaker adaptation respectively. Since in general no prior knowledge of either the channel type, the background noise characteristics or the speaker is available, adaptation is performed using only the test data in an unsupervised mode.

The same tools can be used in acoustic model training in order to compensate for sparse data, as in many cases only limited representative data are available. The basic idea is to use a small amount of representative data to adapt models trained on other large sources of data. Some typical uses are to build gender-dependent, speaker-specific or task-specific models, or to use speaker adaptive training (SAT) to improve performance. When used for model adaption during training it is common to use the true transcription of the data, known as supervised adaptation.

Three commonly used schemes to adapt the parameters of an HMM can be distinguished: Bayesian adaptation [60]; adaptation based on linear transformations [102]; and model composition techniques [48]. Bayesian estimation can be seen as a way to incorporate prior knowledge into the training procedure by adding probabilistic constraints on the model parameters. The HMM parameters are still estimated with the EM algorithm but using maximum a posteriori (MAP) reestimation formulas [60]. This leads to the so-called MAP adaptation technique where constraints on the HMM parameters are estimated based on parameters of an existing model. Speaker-independent acoustic models can serve as seed models for gender adaptation using the gender specific data. MAP adaptation can be used to adapt to any desired condition for which sufficient labeled training data are available. Linear transforms are powerful tools to perform unsupervised speaker and environmental adaptation. Usually these transformations are ML trained and are applied to the HMM Gaussian means, but can also be applied to the Gaussian variance parameters. This ML linear regression (MLLR) technique is very appropriate to unsupervised adaptation because the number of adaptation parameters can be very small. MLLR adaptation can be applied to both the test data and training data. Model composition is mostly used to compensate for additive noise by explicitly modeling the background noise (usually with a single

Gaussian) and combining this model with the clean speech model [47]. This approach has the advantage of directly modeling the noisy channel as opposed to the blind adaptation perform by the MLLR technique when applied to the same problem.

The chosen adaptation method depends on the type of mismatch and on the amount of available adaptation data. The adaptation data may be part of the training data, as in adaptation of acoustic seed models to a new corpus or a subset of the training material (gender, dialect, speaker or acoustic condition specific) or can be the test data (i.e., the data to be transcribed). In the former case supervised adaptation techniques can be applied, as the reference transcription of the adaptation data can be readily available. In the latter case only unsupervised adaptation techniques can be applied.

*IEEE section on Adaptation **JL to merge this in***

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 variability, 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 [47]. 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 [48]. 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 [59, 99]. 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 [156, 185].

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 [36, 102], 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 [177]. 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 [7]. 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 [6]. The ap-

proach 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 [101]) 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 [164].

Decoding

Speech/nonspeech detection

to be done

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 [70, 107, 171], distance-based segmentations [91, 159], and methods based on hypothesis testing [27, 172]. 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 [55]. 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.

Decoding strategies

In this section we discuss the LVCSR decoding problem, which 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 pratice, 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 which uses a dynamic programming algorithm. This basic strategy has been extended to deal with large vocabularies by adding features such as dynamic decoding, multipass search and N-best rescoring.

The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous Viterbi beam search [120] 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 [15, 62], word-dependent phonetic trees [121], forward-backward search [10], N-best rescoring [152], progressive search [119] and simple one-pass dynamic network decoding [124]. An alternative to the frame-synchronous Viterbi beam search is an asynchronous search based on the A* algorithm such as *stack decoding* [14, 136] or the *envelope search* [68].

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 [136] or the one-pass frame synchronous dynamic network decoder [124], is very attractive for real-time applications.

Static decoders require much more memory than dynamic decoders when

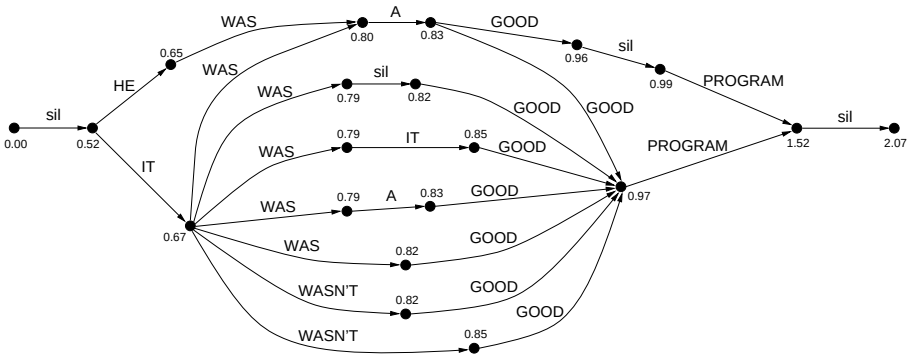

Figure 1.8 Example word lattice generated by a speech recognizer using a bigram language model for a 2.1s utterance. Each graph edge corresponds to a word hypothesis and a time interval (as specified by the time information on the nodes). In this example the word transcription with the highest likelihood is "sil IT WAS A GOOD PROGRAM sil" which happen to be what was said. (Acoustic and language model likelihoods are not given on the figure.)

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[3] corresponding to a recognizer HMM network, substantial reduction of the overall network size can be obtained, enabling static decoding with long span LMs [118]. Evidently, the size of the optimized network remains proportional to the LM size.

Multi-pass decoding is used to progressively add knowledge sources in the decoding process and allows the the complexity of the individual decoding passes to be reduced *from ieee* and often results in a faster overall decoder [122]. 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[4], although some systems use N-best hypotheses (a list of the most likely word sequences with their respectives scores). This approach is not well suited to real-time applications since no hypothesis can be returned until the entire utterance has been processed. *from ieee* However if a small delay is acceptable, then with appropriate synchronization, multipass strategies can be envisioned. Evidently, the first

[3] 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.

[4] Lattices are graphs where nodes correspond to particular frames and where arcs representing word hypothesis have associated acoustic and language model scores.

pass used to generate the initial word lattice must be accurate enough to not introduce lattice errors which are unrecoverable with further processing.

It can sometimes be difficult to add certain knowledge sources into the decoding process especially when they do not fit in the Markovian framework (i.e. short distance dependency modeling). For example, this is the case when trying to use segmental information or to use grammatical information for long term agreement. Such information can be more easily integrated in multipass systems by rescoring the recognizer hypotheses after applying the additional knowledge sources.

Efficiency

** in this section, remove everything which is LIMSI specific, add lantency stuff from ATT **

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 [21]. 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 1 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, lim-

itations 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 [9, 121, 124]. The success of such a single pass approach is highly dependent on the use of efficient pruning strategies associated with a language model lookahead [127, 151]. 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 1 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) [163]. 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 [110, 163], 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 ?? 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 back-off 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 back-off components by increasing the cutoffs on the n-gram counts, thus reducing significantly the LM size. More elaborate n-gram pruning have also been proposed [158, 161] 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 [140].

Confidence Measures

expand idea of lattice based CM, plus consensus decoding confidence *

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 [23, 63, 160, 173, 174]. 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 [53]. 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 [173]. 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 [63], a generalized additive model [160], or a neural-network [174]. 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.

Indicative Performance levels

** may want subsections here for different tasks** **dictation systems, spoken language dialog systems, and transcription systems for audio indexation**

add section on audio indexing??

This section provides somes indicative measures of recognizer performance for a few LVCSR tasks, but does not attempt to be exhaustive. Essentially all of todays state-of-the-art systems make use of the statistical modeling presented in this chapter. Speech recognition technology has advanced greatly over the last decade. These advances can be clearly seen in the context of DARPA supported benchmark evaluations. This framework, known in the community as the DARPA evaluation paradigm, has provided the training materials (transcribed audio and textual corpora for training acoustic and language models), test data and a common evaluation framework. In recent years the data have been provided by the Linguistics Data Consortium (LDC) and the evaluations organized by the National Institute of Standards and Technology (NIST) in collaboration with representatives from the participating sites and other government agencies. It is widely acknowledged that the performance of a speech recognizer is strongly dependent upon the task, which in turn is linked to the type of user, speaking style, environmental conditions etc.

The commonly used metric for speech recognition performance, the "word error" rate, is a measure of the average number of errors taking into account three error types with respect to a reference transcription: *substitutions* (the reference word is replaced by another word), *insertions* (a word is hypothesized that was not in the reference) and *deletions* (a word in the reference transcription is missed). The word error rate is defined as

$$\frac{\#\text{subs}+\#\text{ins}+\#\text{del}}{\#\text{ reference words}}$$

, and is generally computed aligning the reference and hypothesized transcriptions using a dynamic programming algorithm, where costs are associated with the different error types. Given this definition the word error can be more than 100%.

While this chapter addresses speech transcription (i.e., going from the audio signal to words), it should be kept in mind that additional information can be extracted from the audio signal. Extraction of some of this so-called "metadata", is discussed in Chapters **schwartz and **Allen. The metadata can be of an acoustic nature (speaker and gender information [96], audio type information [51, 157]) or linguistic nature (case-sensitive texts, punctuation, named entities (names of persons, places, organizations), topics, or other semantic tags. The same HMM-based probabilistic framework has been used to assign tags [115, 170, 178]. 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.

Dictation is the most obvious automatic speech recognition task, and has a long history of research and product development, resulting in the low-cost, off-the-shelf systems for a variety of platforms and languages. While from the technological viewpoint, dictation is usually thought of as a "simple"

transformation from speech to text, this view overlooks a variety of formatting and integration issues which are important for usability. Perhaps the most notable characteristic of the dictation task is that the speech data is produced with the explicit goal of being transcribed by a machine. The speech data in a dictation session comes from a single speaker and is recorded with a controlled signal acquisition setup. The linguistic content is usually somewhat limited and the word stream is quite close to the written form.

Although benchmarks of commercial dictation systems are not publicly available, 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. This close relation between system development and evaluation, which has been referred to as "assessment driven technology development" had led to large performance improvements in spite of increasing task difficulty. For read speech tasks, the state-of-the-art in speaker-independent continuous speech recognition is exemplified by the 1995/1996 benchmark tests on North American Business News task [131, 132]. 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.[5] 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 [90]. 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.

maybe this should be later? 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 [181], which aimed to assess language-dependent issues in multilingual recognizer evaluation, demonstrated that the same recognition technology and evaluation methodology used for American

[5] With the exception of the telephone recordings, the speakers were allowed to repeat their recording if unsatisfied with it. [?]

English could be successfully applied to a dictation task in British English, French and German.

The speech recognizer is often considered a critical component of spoken dialog systems, which aim to enable vocal access to stored information. In order to provide user-friendly interaction with a machine, it is necessary to be able to recognize naturally spoken spontaneous utterances from unknown speakers. In general each user interacts only briefly with the machine, so there is very little data available for model adaptation. Telephone services are a natural area for spoken dialog systems as the only means of interaction with the machine are via voice 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 of particular importance in the context of natural, mixed-initiative dialogs. Growing in popularity are information kiosks [52] and multimedia web interfaces, in which different modalities (tactile and audio) can be used for input and output. The speech recognition component of dialog systems are typically faced with more challenging acoustic conditions than for dictation tasks, being subject to channel distortions, varied handsets and noisy background conditions. The capability of the user to interrupt the machine is often considered as crucial for usability.

In contrast to dictation applications where it is relatively straight-forward to obtain large written corpora for language modeling, for dialog systems it is usually necessary to collect application-specific data, which can represent a significant portion of the development effort [97]. 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.

The most widely known efforts in evaluation of SLDSs are the DARPA ATIS task [72, 109, 138], the German national Verbmobil project [169] and the EC Language Engineering projects [111, 112]. 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. It should be noted that reporting word error rates can be somewhat misleading, since all differences between the exact orthographic form of the query and the recognizer output are counted as errors, and some of recognition errors (such as gender or plurals) are not important for understanding. A more appropriate measure could be the error rate on meaningful words or concepts used in later processing stages. For instance, in the (DARPA ATIS benchmark tests [130, 131]) the understanding error based on the spoken input was not much larger than the natural language understanding error obtained using manual orthographic transcriptions. In the case of multimodal systems,

the effectiveness of speech must be assessed in coordination with the other modalities.

A more recent application area is the transcription and indexation of general audio data, such as radio and television broadcasts[6], or meetings and teleconferences, and any kind of audio data mining. Several characteristics of this type of audio data can be noted. Firstly, 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. Secondly, 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. Thirdly, the signal capture and background environment can be only more or less controlled.

Substantially higher word error rates, above 30-40% have been reported for the transcription of telephone conversational speech using the Switchboard and multilinugal Callhome (Spanish, Arabic, Mandarin, Japanese, German) corpora. While most of the results given here are for American English, somewhat comparable results have been reported by various sites for other languages including French, German and British English.

Over the last few years there has been increasing interest in the transcription of radio and television broadcasts, often referred to as "found speech." This is a major step for the community in that the test data is taken from a real task, as opposed to consisting of data recorded for evaluation purposes. The transcription of such broadcasts presents new challenges as the signal is one continuous audio stream that contains segments of different acoustic and linguistic natures. Systems trained on 150h of acoustic data and 200 M words of commercial transcripts achieve word error rates around 20% on unrestricted broadcast news data. The peformance on studio quality speech from announcers is comparable to that obtained on WSJ read speech data.

———******taken from ieee**********———

to be substantially reduced *

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

[6]The earliest work in this area that we are aware of is the NSF INFORMEDIA project [71] under the Digital Libraries News-on-Demand action line. A special section of the Communications of the ACM was recently devoted to this topic [114].

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 ??, 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 [51], 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 [133, 134, 135]. 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 [44]. 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 [22, 81, 86, 126, 135] 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 recog-

nition task using the Switchboard [65] and multilingual Callhome (Spanish, Arabic, Mandarin, Japanese, German) corpora. The word rates reported for this data, on the order of 30-40% [182], 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 [50, 168]. The average word error measured on a representative 10 hour subset of this data was around 20% for state-of-the-art systems [50]. 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.

———******end from ieee**********————

Language Dependencies & Portability

Statistically-based speech recognition technology has been successfully employed for a variety of tasks and languages. The porting of a recognition system to a new task or another language requires the availability of sufficient amounts of transcribed training data and substantial effort is involved to construct the acoustic and language models, and to develop the recognition lexicon. 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). However, 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 [64]. 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 [92] 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.

Substantial effort may be 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.

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.

Language portability is particular important for audio indexation tasks, where a a characteristic 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 multi-lingual indexation and retrieval. Multilinguality is thus of

particular interest for media watch applications, where news may first break in another country or language. **remove limsi??** The LIMSI American English broadcast news transcription system has been ported to six other languages. 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. (see table)

Language	*#phones*	*Lexicon size*	*coverage*	*N-gram perplexity*	*Test %Werr*
English	48	65k	99.4%	140	20
French	37	65k	98.8%	98	23
German	51	65k	96.5%	213	25
Mandarin	39	40k+5k chars	99.7%	190	20
Spanish	27	65k	94.3%	159	20
Portuguese	39	65k	94.0%	154*	40
Arabic	40	65k	90.5%	160*	20

Figure 1.8 Some language characteristics. For each language are specified: the number of phones used to represent lexical pronunciations, the approximate vocabulary size in words (characters for Mandarin) and lexical coverage (of the test data), the test data perplexity with a 4-gram language models *(3-gram for Portuguese and Arabic), duration and the word/character error rates. For Arabic the vocabulary and language model are vowelized, however the word error rate does not include vowel or gemination errors.

Porting a recognizer to another language necessitates modifying those system components which incorporate language-dependent knowledge sources such as the phone set, the recognition lexicon, phonological rules and the language model. Other considerations are the 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. There are two predominant approaches taken to bootstrapping the acoustic models for another language. The first is to use acoustic models from an existing recognizer and a pronunciation dictionary to segment manually annotated training data for the target language. If recognizers for several languages are available, the seed models can be selected by taking the closest model in one of the available language-specific sets. An alternative approach is to use a set of global acoustic models, that cover a wide number of phonemes [?]. This approach offers the advantage of being able to use multilingual acoustic models to provide additional training data, which is particularly interesting when only very limited amounts of data ($<$ 10 hours) for the target language are available.

There are some notable language specificities. The number of phones use in the recognition lexicon is to range from 25 for Spanish to 51 for German (see Table 1.1). The Mandarin phone set distinguishes 3 tones, which are

associated with the vowels. If the tone distinctions are taken into account the Mandarin phone set differentiates 61 units. For most of the languages it is reasonably straightforward to generate pronunciations (and even some predictable variants) using grapheme-to-phoneme rules. The automatically generated pronunciations can optionally be manually verified. A notable exception is the English language for which most of the pronunciations have been manually derived. Another important language characteristic is the lexical variety. The agglutination and case declension in German results in a significantly larger OOV rate for a fixed size lexicon. French, Spanish and Portuguese all have gender and number agreement which expands the lexical variety, which for French also leads to a high homophone rate, particularly for verb forms. The Mandarin language also poses the problem of word segmentation, but this is offset by the opportunity to eliminate OOVs by including all characters in the recognition word list [?]. The Arabic language also is agglutinative, but a larger challenge is to handle to lack of vowelization in written texts. This is compounded by a wide variety of Arabic dialects, many of which do not have a written form.

At LIMSI broadcast news transcription systems have been developed for the American English, French, German, Mandarin, Spanish, Arabic and Portuguese languages. To give an idea of the resources used in developing these systems, there are roughly 200 hours of transcribed audio data for American English, about 50 hours for French and Arabic, 20 hours for German, Mandarin and Spanish, with 3.5 hours for Portuguese. The data come from a variety of radio and television sources. Obtaining appropriate language model training data is also difficult. While newspaper and newswire texts are becoming widely available in many languages, these texts are quite different than transcriptions of spoken language. There are also significantly more language model training texts available for American English (over 1 billions words including 240 million words 10k hours of commercially produced transcripts). For the other languages there are on the order of 200-300 millions words of texts, with the exception of Portuguese where only 70 millions words are available. It should be noted that French is the only language other than American English for which commercially produced transcripts are available (20 million words).

Some of the system characteristics are shown in Table 1.1, along with indicative recognition performance rates. The word error rate on unrestricted American English broadcast news data is about 20% [?, ?]. The transcription systems for French and German have comparable error rates for news broadcasts [?]. The character error rate for Mandarin is also about 20% [?]. Based on our experience, it appears that with appropriately trained models, recognizer performance is more dependent upon the type and source of data, than on the language. For example, documentaries are particularly challenging to transcribe, as the audio quality is often not very high, and there is a large proportion of voice over.

With today's technology, the adaptation of a recognition system to a dif-

ferent task or language requires the availability of sufficient amounts of transcribed training data. Obtaining such data is usually an expensive process in terms of manpower and time. Recent work has focused on reducing this development cost [?].

Standard HMM training requires an alignment between the audio signal and the phone models, which usually relies on an orthographic transcription of the speech data and a good phonemic lexicon. The orthographic transcription is usually considered as ground truth, that is the word sequence should by hypothesized by the speech recognizer when confronted with the same speech segment. One can imagine training acoustic models in a less supervised manner. Any available related linguistic information about the audio sample can be used in place of the manual transcriptions required for alignment, by incorporating this information in a language model, which can be used to produce the most likely word transcription given the current models. An iterative procedure can successively refine the models and the transcription.

One approach is to use existing recognizer components (developed for other tasks or languages) to automatically transcribe task-specific training data. Although in the beginning the error rate on new data is likely to be rather high, this speech data can be used to re-train a recognition system. If carried out in an iterative manner, the speech corpus can be cumulatively extended over time *without* direct manual transcription. This approach has been investigated in [86, ?, ?].

There are certain audio sources such as radio and television news broadcasts, that can provide an essentially unlimited supply of acoustic training data. However, for the vast majority of audio data sources there are no corresponding accurate word transcriptions. Some of these sources, in particular, the main American television channels also broadcast manually derived closed-captions. The closed-captions are a close, but inexact transcriptions of what is spoken, and are only coarsely time-aligned with the audio signal. Manual transcripts are also available for certain radio broadcasts. There also exist other sources of information with different levels of completeness such as approximate transcriptions or summaries, which can be used to provide some supervision.

Experiments exploring lightly supervised acoustic model training were carried out using unannotated audio data containing over 500 hours of BN audio data [?]. First the recognition performance as a function of the available acoustic training data was assessed. With 200 hours of manually annotated acoustic training data (the standard Hub4 training data), a word error rate of 18.0% was obtained. Reducing the training data by a factor of two increases the word error rate to 19.1%, and by a factor of 4 to 20.7%. With only 1 hour of training data, the word error rate is 33.3%. A set of experiments investigated the impact of different levels of supervision via the language model training materials. Language models were estimated using various combinations of the text sources, from the same epoch as the audio data or predating the period. Since newspaper and newswire sources have only an indirect corre-

spondence with the audio data, they provide less supervision than the closed captions and commercially generated transcripts [?]. While all of the language models provided adequate supervision for the procedure to converge, and those that included commercially produced transcripts in the training material performed slightly better. It was found that somewhat comparable acoustic models could be estimated on 400 hours automatically annotated BN data and 200 hours of carefully annotated data.

This unsupervised approach was used to develop acoustic models for the Portuguese language for which substantially less manually transcribed data are available. Initial acoustic model trained on the 3.5 hours of available data were used to transcribe 30 hours of Portuguese TV broadcasts. These acoustic models had a word error rate of 42.6%. By training on the 30 hours of data using the automatic transcripts the word error was reduced to 39.1%. This preliminary experiment supports the feasibility of lightly supervised and unsupervised acoustic model training.

Improving Genericity

In the context of the EC CORETEX project, research is underway to improve the genercity of speech recognition technology, by improving the basic technolgoy and exploring rapid adaptation methods which start with the initial robust generic system and enhance performance on particular tasks. To this extent, cross task recognition experiments have been reported where models from one task are used as a starting point for other tasks [?, ?, ?, ?, ?, ?]. In [?] broadcast news (BN) [?] acoustic and language models to decode the test data for three other tasks (TI-digits [?], ATIS [?] and WSJ [?]). For TI-digits and ATIS the word error rate increase was shown to be primarily due to a linguistic mismatch since using task-specific language models greatly reduces the error rate. For spontaneous WSJ dictation the BN models out-performed task-specific models trained on read speech data, which can be attributed to a better modelization of spontaneous speech effects (such as breath and filler words).

Methods to improve genericity of the models via multi-source training have been investigated. Multi-source training can be carried out in a variety of ways – by pooling data, by interpolating models or via single or multi-step model adaptation. The aim of multi-source training is to obtain generic models which are comparable in performance to the respective task-dependent models for all tasks under consideration. Compared to the results obtained with task-dependent acoustic models, both data pooling and sequential adaptation schemes led to better performance for ATIS and WSJ read, with slight degradations for BN and TI-digits [?].

In [?] cross-task porting experiments are reported for porting from an Italian broadcast news speech recognition system to two spoken dialogue domains. Supervised adaptation was shown to recover about 60% of the WER gap between the broadcast news acoustic models and the task-specific acoustic

models. Language model adaptation using just 30 minutes of transcriptions was found to reduce the gap in perplexity between the broadcast news and task-dependent language models by 90%. It was also observed that the out-of-vocabulary rates for the task-specific language models are 3 to 5 times higher than the best adapted models, due to the relatively limited amount of task-specific data and the wide coverage of the broadcast news domain.

Techniques for large-scale discriminative training of the acoustic models of speech recognition systems using the maximum mutual information estimation (MMIE) criterion in place of conventional maximum likelihood estimation (MLE) have studied and it has been demonstrated that MMIE-based systems can lead to sizable reductions in word error rate on the transcription of conversational telephone speech [?]. Experiments on discriminative training for cross-task genericity have made use of recognition systems trained on the low-noise North American Business News corpus of read newspaper texts and tested on television and radio Broadcast News data. These experiments showed that MMIE-trained models could indeed provide improved cross-task performance [?].

Perspectives

robustness

improving model genericity and portability

Unsupervised training for AM and LMs

Despite the advances in technology witnessed over the last decode recognition performance remains highly dependent upon the task, the talker, speaking style, recording and environmental conditions etc.

Despite recent progress, automatic speech recognition performance remains far from human performance [34, 41, 106, 165], with differences in the range of 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 [43] 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 [131]. 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. Reduc-

ing 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 dictation with typing is given in [84].

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 [86] 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 [64]. 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 [93] 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 [145] is a promising research direction to improve the usability of multimodal kiosks.

Despite the numerous advances made over the last decade, speech recognition is far from a solved problem. Much of the recent progress in LVCSR has been made fostered by the availability of large corpora for training and testing speech recognition and understanding technology. However, 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. Re-

cent efforts have been directed at developing generic recognition models and the use of unannotated data for training purposes, in an aim to reduce the reliance on manually annotated training corpora.

It has often been observed that there is a large difference in recognition performance for the same system between the best and worst speakers. Unsupervised adaption techniques do not necessarily reduce this difference, in fact, often they improve performance on good speakers more than on bad ones. Interspeaker differences are not only at the acoustic level, but also the phonological and word levels. Today's modeling techniques are not able to take into account speaker-specific lexical and phonological choices.

A wide range of potential applications can be envisioned based on transcriptions of broadcast data, particularly in light of the recent explosion of such media, which required automated processing for indexation and retrieval (Chapters 29, 30 and 32). Related spoken language technologies, such as speaker and language identification, which rely on the same modeling techniques, are clearly of interest for automated processing of large multilingual corpora. Important future research will address keeping vocabulary up-to-date, language model adaptation, automatic topic detection and labeling, and enriched transcriptions providing annotations for speaker turns, language, acoustic conditions, etc.

References

[1] **Read-Hill, R. E.**, *Physical Metallurgy Principles,* D. van Nostrand, New York, 1973, Chap. 13.

[2] **Porter, D. A. and Easterling, K. E.**, *Phase Transformations in Metals and Alloys,* Chapman and Hall, London, 1992.

[3] **Christian, J. W.**, *The Theory of Transformations in Metals and Alloys,* Pergamon, Oxford, 1975, Chap. 10, 11 and 12.

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

[5] **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.

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

[7] **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.

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

[9] **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.

[10] **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.

[11] **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.

[12] **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.

[13] **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.

[14] **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.

[15] **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.

[16] **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.

[17] Baum, L.E., T. Petrie, G. Soules, and N. Weiss. 1970. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". *Ann. Math. Stat.*. **41**. 164-171.

[18] **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.

[19] **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.

[20] **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.

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

[22] **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.

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

[24] **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.

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

[26] **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.

[27] **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.

[28] **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.

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

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

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

[32] 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.

[33] Dempster, A.P., M.M. Laird and D.B. Rubin. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society Series B (methodological).* **39**: 1-38.

[34] 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.

[35] 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.

[36] 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.

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

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

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

[40] 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.

[41] 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.

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

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] I.J. Good, "The Population Frequencies of Species and the Estimation of Population Parameters," *Biomterika*, **40**(3/4):237-264, 1953.

[68] 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.

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

[70] 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.

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

[72] 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.

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

[74] 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.

[75] M.J. Hunt. 1996. "Signal Representation," Chapter 1.3 of the State of the Art in Human Language Technology, (Cole et al, eds.) (http://www.cse.ogi.edu/CSLU/HLTsurvey/ch1node2.html)

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

[77] 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.

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

[79] F. Jelinek, *Statistical Methods for Speech Recognition,* Cambridge: MIT Press, 1997.

[80] 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.

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

[82] 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.

[83] Juang, B.-H. 1985. "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT&T Technical Journal*, **64**(6).

[84] 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.

[85] 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.

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

[87] 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.

[88] R. Kneser and H. Ney, *Improved Clustering Techniques for Class-Based Statistical Language Modelling, Proc. Eurospeech'93*, pp. 973-976, Berlin, **** 1993.

[89] 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.

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

[91] 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.

[92] 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.

[93] 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.

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

[95] 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.

[96] 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.

[97] 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.

[98] 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*.

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

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

[101] 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.

[102] 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.

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

[104] 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.

[105] Liporace, L. R. 1982. "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Transactions on Information Theory*, **IT-28**(5). 729-734.

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

[107] 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.

[108] 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.

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

[110] 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.

[111] 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.

[112] J.J. Mariani and L.F. Lamel, *An overview of EU programs related to conversational/interactive systems, Proc.* DARPA *Broadcast News Tran-*

scription & Understanding Workshop, pp. 247-253, Landsdowne,VA, February 1998.

[113] S. Martin, J. Liermann and H. Ney (1995). *Algorithms for Bigram and Trigram Clustering, Proc. Eurospeech'95*, pp. 1253-1256, Madrid, *** 1995.

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

[115] 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.

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

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

[118] 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.

[119] 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.

[120] 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.

[121] 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.

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

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

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

[138] 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.

[139] L.R. Rabiner, and B.H. Juang, 1986. "An Introduction to Hidden Markov Models. *IEEE Acoustics Speech and Signal Processing Magazine.* **ASSP-3**(1). 4-16. January.

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

[141] 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.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

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

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

[148] R. Rosenfeld, *Two Decades of Statistical Language Modeling: Where Do We Go From Here?, Proceedings of the IEEE*, **88**(8), 1270-1278, August 2000.

[149] 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.

[150] 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.

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

[152] 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.

[153] 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.

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

[155] 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.

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

[157] 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.

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

[159] 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.

[160] M. Siu and H. Gish, *Evaluation of word confidence for speech recognition systems*, *Computer Speech & Language*, **13**(4), pp. 299-318, October 1999.

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

[162] 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.

[163] 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.

[164] 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.

[165] 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.

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

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

[168] 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.

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

[170] 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.

[171] 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.

[172] 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.

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

[174] 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.

[175] I.H. Witten and T.C. Bell, "The Zero Frequency problem: Estimating the problems of Novel Events in Adaptive tex Compression," *Proc. IEEE Trans. on Information Theory,* **37**(), 1085-1094, July 1991.

[176] 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.

[177] 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.

[178] 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.*

[179] 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.*

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

[181] 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.*

[182] 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.*

[183] 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.*

[184] *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.*

[185] *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.*

[186] *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.