SPEECH RECOGNITION

Lori Lamel and Jean-Luc Gauvain

Abstract

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 speech recognition, that is acoustic-phonetic modeling, lexical representation, language modeling, decoding and model adaptation. The focus is on methods used in state-of-the-art speaker-independent, large vocabulary continuous speech recognition (LVCSR). Some of the technology advances over the last decade are highlighted. Primary application areas for such technology initially addressed dictation tasks and interactive systems for limited domain information access (usually referred to as spoken language dialog systems. The last decade has witnessed a wider coverage of languages, as well as growing interest in transcription systems for information archival and retrieval, media monitoring, automatic subtitling and speech analytics. Some outstanding issues and directions of future research are discussed.

1 Overview

Speech recognition is principally concerned with the problem of transcribing the speech signal as a sequence of words. Today's best performing systems use statistical models (Chapter 19) of speech. From this point of view, speech is assumed to be generated by a **language model** which provides estimates of $\Pr(w)$ for all word strings w independently of the observed signal, and an **acoustic model** encoding the message w in the signal x, which is represented by a probability density function $f(x|w)$. The goal of speech recognition is to find 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 $\Pr(w)f(x|w)$.

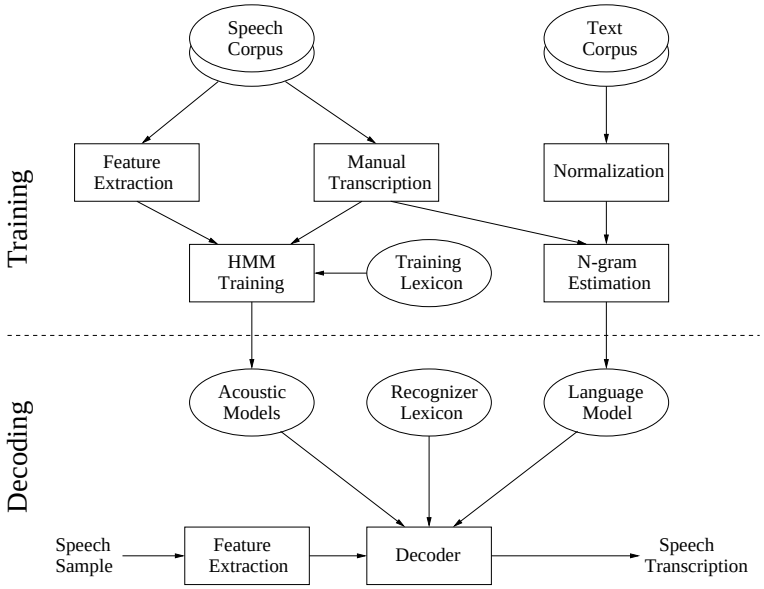

Figure 1: System diagram of a generic speech recognizer based using statistical models, including training and decoding processes.

The principles on which these systems are based have been known for many years now, and include the application of information theory to speech recognition (Bahl et al. 1976; Jelinek 1976), the use of a spectral representation of the speech signal (Dreyfus-Graf 1949; Dudley and Balashek 1958), the use of dynamic programming for decoding (Vintsyuk 1968), and the use of **context-dependent** acoustic models (Schwartz et al. 1984). 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 (Chapter 24), 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 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. 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 remaining sections of this chapter are devoted to discussing these main components.

2 Acoustic parameterization and 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 (HMM) for acoustic modeling, which consists of modeling the probability density function of a sequence of acoustic feature vectors. In this section common parameterizations are described, followed by a discussion of acoustic model estimation and adaptation.

2.1 Acoustic feature analysis

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 (Hunt 1996). 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 HMM 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 (Davis and Mermelstein 1980) or with a Perceptual Linear Prediction (PLP) analysis (Hermansky

1990). 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 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. The set of cepstral coefficients associated with a windowed portion of the signal is referred to as a **frame** or a **parameter vector**. Cepstral mean removal (subtraction of the mean from all input frames) 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). 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. Over the last decade there has been growing in interest in capturing longer term dynamics of speech than of the standard cepstral features. A variety of techniques have been proposed from simple concatenation of sequential frames to the use of TempoRAl Patterns (TRAPs) (Hermansky and Sharma, 1998), In all cases the wider context results in a larger number of parameters that consequently need to be reduced. Discriminative classifiers such as Multi Layer Perceptrons (MLP) are efficient methods for discriminative feature estimation. Over the years, several groups have developed mature techniques for extracting probabilistic

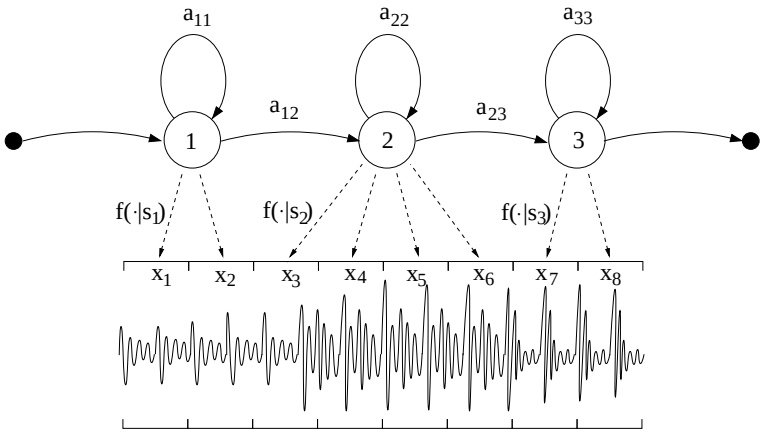

Figure 2: A typical 3-state phone HMM with no skip state (top) which generates feature vectors ($x_1 ... x_n$) representing speech segments.

MLP features and incorporating them in speech-to-text systems (Zhu et al., 2005; Stolcke et al., 2006). While probabilistic features have not been shown to consistently outperform cepstral features in LVCSR, being complementary they have been shown to significantly improve performance when used together (Fousek et al., 2008).

2.2 Acoustic models

Hidden Markov models are widely used to model the sequences of acoustic feature vectors (Rabiner and Juang 1986). 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. The most widely used elementary acoustic units in LVCSR systems are phone-based [1], where each **phone**[2] is represented by a Markov chain with a small

[1] For some languages, most notably tonal languages such as Chinese longer units corresponding to syllables or demisyllables (also called onsets and offsets or initials and finals) have been explored. While the use of larger units remains relatively limited to phone units, they may better capture tone information and may be well-suited to casual speaking styles.

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

number of 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 $f(\mathbf{x}, \mathbf{s}|\lambda)$ of the observed signal $\mathbf{x} = (x_1, ..., x_T)$ and the unobserved state sequence $\mathbf{s} = (s_0, ..., s_T)$,

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

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 2 shows a 3-state HMM with the associated transition probabilities and observation PDFs.

Phone based models offer the advantage that recognition lexicons can be described using the elementary units of the given language, and thus benefit 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 (Bahl et al. 1988). 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). 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 (see section 5). 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). The model states are often clustered so as to reduce the model size, resulting in what are referred to as "tied-state" models.

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 (Dempster et al. 1977) 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 (Baum et al. 1970; Liporace 1982; Juang 1985). 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, which requires automatic identification of the gender.

2.3 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 (Gauvain and Lee 1994); adaptation based on linear transformations (Leggetter and Woodland 1995); and model

composition techniques (Gales and Young 1995). 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 (Gauvain and Lee 1994). 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. 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.

3 Lexical and pronunciation modeling

The **lexicon** is the link between the acoustic-level representation and the word sequence output by the speech recognizer. Lexical design entails two

main parts: 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.

The recognition vocabulary is usually selected to maximize lexical coverage for a given size lexicon. Since on average, each out-of-vocabulary (OOV) word causes more than a single error (usually between 1.5 and 2 errors), it is important to judiciously select the recognition vocabulary. Word list selection is discussed in Section 4. 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, 49 for German, 35 for French, and 26 for Spanish. In generating pronunciation baseforms, most lexicons include standard pronunciations and do not explicitly represent **allophones**. This representation is chosen as most allophonic variants can be predicted by rules, and their use is optional. More importantly, there often is a continuum between different allophones of a given phoneme and the decision as to which occurred in any given utterance is subjective. By using a phonemic representation, no hard decision is imposed, and it is left to the acoustic models to represent the observed variants in the training data. While pronunciation lexicons are usually (at least partially) created manually, several approaches to automatically learn and generate word pronunciations have been investigated (Cohen 1989; Riley and Ljojle 1996).

There are a variety of words for which frequent alternative pronunciation variants are observed, and these variants are not due to allophonic differences such as the suffix *-ization* which can be pronounced with a diphthong (/ɑ^i/) or a schwa (/ə/). Alternate pronunciations are also needed for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as *excuse, record, produce*. Some common 3 syllable words such as *interest* and *company* are often pronounced with only 2 syllables. Figure 3 shows two examples of the word *interest* by different speakers reading the same text prompt: "*In reaction to the news, interest rates plunged...*". The pronunciations are those chosen by the recognizer during segmentation using forced alignment. In the example on the left,

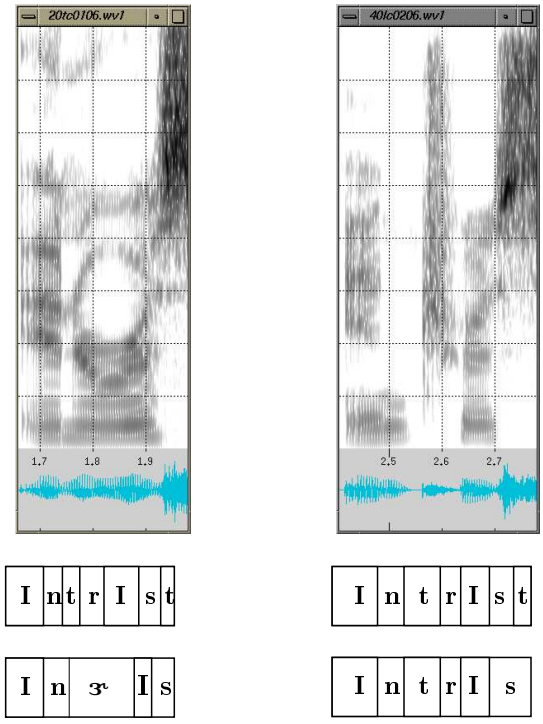

Figure 3: Spectrograms of the word *interest* with pronunciation variants: /ɪnɝɪs/ (left) and /ɪntrɪs/(right) taken from the WSJ corpus (sentences 20tc0106, 40lc0206). The grid is 100ms by 1 kHz. Segmentation of these utterances with a single pronunciation of *interest* /ɪntrɪst/ (middle) and with multiple variants /ɪntrɪst/ /ɪntrɪs/ /ɪnɝɪs/ (bottom). The /I/ and /t/ segments are light and dark grey respectively.

the /t/ is deleted, and the /n/ is produced as a nasal flap. In the example on the right, the speaker said the word with 2 syllables, without the optional vowel and producing a /tr/ cluster. Segmenting the training data without pronunciation variants is illustrated in the middle. Whereas no /t/ was observed in the first pronunciation example two /t/ segments had been aligned. An optimal alignment with a pronunciation dictionary including all required variants is shown on the bottom. Better alignment results in more accurate acoustic phone models. Careful lexical design improves speech recognition performance.

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 for long words such as *positioning* (/pəzɪʃəniŋ/ or /pəzɪʃniŋ/), can be included in the lexicon allowing 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 "did you" pronounced as "dija" or "going to" pronounced as "gonna". Alternatively, such fluent speech effects can be modeled using phonological rules (Oshika et al. 1975). 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. Most of todays state-of-the-art systems include pronunciation variants in the dictionary, associating pronunciation probabilities with the variants (Bourlard et al., eds. 1999; Fosler-Lussier et al., eds., 2005).

As speech recognition research has moved from read speech to spontaneous and conversational speech styles, 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).

4 Language modeling

Language models (LMs) 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 $(w_1, w_2, ..., w_k)$ can be approximated by $\prod_{i=1}^{k} \Pr(w_i|w_{i-n+1}, ..., w_{i-2}, w_{i-1})$, therefore reducing the word history to the preceding $n-1$ words. A **back-off** mechanism is generally

used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved word sequences (Katz 1987). 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.

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. 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.

A common motivation for normalization in all languages 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 *Gates, gates* or *Green, green.* In the French *Le Monde* corpus, capitalization of proper names is distinctive with different lexical entries for *Pierre, pierre* or *Roman, roman.*

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 are typically expanded to approximate the spoken form ($150 → one hundred and fifty dollars). 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 → A. B. C. D.). Another example is the treatment of numbers in German, where decompounding can be used in order to increase lexical coverage. The date 1991 which in standard German is written as *neunzehnhunderteinundneunzig* can be represented by word sequence *neunzehn hundert ein und neunzig*. 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 language models can be obtained using texts transformed to be closer to the observed reading style, where the transformation rules and corresponding probabilities are automatically derived by aligning prompt texts with the transcriptions of the acoustic data. For example, the 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 *million dollars* are replaced with simply *million*.

In practice, the selection of words is done so as to minimize the system's OOV rate by including the most useful words. By useful we mean that the words are expected as an input to the recognizer, but also that the LM can be trained given the available text corpora. In order to meet the latter condition, it is common to choose the N most frequent words in the training data. This criterion does not, however, guaranty the usefulness of the lexicon, since no consideration of the expected input is made. Therefore it is common practice to use a set of additional development data to select a word list adapted to the expected test conditions.

There is the sometimes conflicting need for sufficient amounts of text data to estimate LM parameters and assuring that the data is representative of the task. It is also common that different types of LM training material are available in differing quantities. One easy way to combine training material from different sources is to train a language model per source and to interpolate them. The interpolation weights can be directly estimated on some development data with the EM algorithm. An alternative is to simply merge the n-gram counts and train a single language model on these counts. If some data sources are more representative than others for the task, the n-gram counts can be empirically weighted to minimize the perplexity on a set

of development data. While this can be effective, it has to be done by trial and error and cannot easily be optimized. In addition, weighting the n-gram counts can pose problems in properly estimating the backoff coefficients. For these reasons the language models in most of todays state-of-the-art systems are obtained via the interpolation methods, which can also allow for task adaptation by simply modifying the interpolation coefficients (Chen et al,. 2004; Liu et al, 2008).

The relevance of a language model is usually measured in terms of test set perplexity defined as Px= $\Pr(\text{text}|\text{LM})^{-\frac{1}{n}}$, where n is the number of words in the text. The perplexity is 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.

5 Decoding

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 practice, however, it is common to search for the most likely HMM state sequence, i.e. the best path through a trellis (the search space) where each node associates an HMM state with given time. Since it is often prohibitive to exhaustively search for the best path, techniques have been developed to reduce the computational load by limiting the search to a small part of the search space. Even for research purposes, where real-time recognition is not needed there is a limit on computing resources (memory and CPU time) above which the development process becomes too costly. The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous Viterbi beam search 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.

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 is very attractive

for real-time applications. 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. 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 graphs, although some systems use N-best hypotheses (a list of the most likely word sequences with their respective scores). This approach is not well suited to real-time applications since no hypothesis can be returned until the entire utterance has been processed.

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.

Mangu, Brill and Stolcke (2000) proposed the technigue of confusion network decoding (also called consensus decoding) which minimizes an approximate WER as opposed to MAP decoding which minimizes the sentence error rate (SER). This technique has since been adopted in most state-of-the-art systems, resulting in lower WERs and better confidence scores. Confidence scores are a measure of the reliability of the recognition hypotheses, and give an estimate of the word error rate. For example, an average confidence of 0.9 will correspond to a word error rate of 10% if deletions are ignored. Jiang (2004) provides an overview of confidence measures for speech recognition, commenting on the capacity and limitations of the techniques.

6 State-of-the-Art Performance

The last decade has seen large performance improvements in speech recognition, particularly for large vocabulary, speaker-independent, continuous speech. This progress has been substantially aided by the availability of large speech and text corpora and by significant inreases in computer processing

capabilities which have facilitated the implementation of more complex models and algorithms.[3] In this section we provide some illustrative results for different LVCSR tasks, but make no attempt to be exhaustive.

The commonly used metric for speech recognition performance is the "word error" rate (Chapter 22), which is a measure of the average number of errors taking into account three error types with respect to a reference transcription: *substitutions* (one word is replaced by another word), *insertions* (a word is hypothesized that was not in the reference) and *deletions* (a word is missed). The word error rate is defined as $\frac{\#\text{subs}+\#\text{ins}+\#\text{del}}{\#\text{ reference words}}$, and is typically computed after a dynamic programming alignment of the reference and hypothesized transcriptions. Note that given this definition the word error can be more than 100%.

Three types of tasks can be considered: small vocabulary tasks, such as isolated command words, digits or digit strings; medium-size (1000-3000 words) vocabulary tasks such as are typically found in spoken dialog systems (Chapter 34); and large vocabulary tasks (typically 65k words). Another dimension is the speaking style which can be read, prepared, spontaneous or conversational. Very low error rates have been reported for small vocabulary tasks, below 1% for digit strings, which has led to some commercial products most notably in the telecommunications domain. Early benchmark evaluations focused on read speech tasks: the state-of-the-art in speaker-independent continuous speech recognition in 1992 is exemplified by the Resource Management task (1000 word vocabulary, word-pair grammar, 4h acoustic training data) with a word error rate of 3%. In 1995, on read newspaper texts (the Wall Street Journal task, 160h acoustic training data and 400 M words of language model texts) word error rates around 8% were obtained using a 65k word vocabulary. The word errors roughly doubled for speech in the presence of noise, or on texts dictated by journalists. The maturity of the technology led to the commercialization of speaker-dependent

[3]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. The data have been generally 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.

continuous speech dictation systems for which comparable benchmarks are not publicly available.

Over the last decade the research has focused on "found speech", originating with the transcription of radio and television broadcasts and moving to any audio found on the Internet (podcasts). This was 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 varied presents new challenges as the signal is one continuous audio stream that contains segments of different acoustic and linguistic natures. Today well-trained transcription systems for broadcast data have been developed for at least 15 languages, achieving word error rates on the order of under 20% on unrestricted broadcast news data. The performance on studio quality speech from announcers is often comparable to that obtained on WSJ read speech data.

Word error rates of under 20% have been reported for the transcription of conversational telephone speech (CTS) in English using the Switchboard corpus, with substantially higher WERs (30-40%) on the multilinugal Callhome (Spanish, Arabic, Mandarin, Japanese, German) data. A wide range of word error rates have been reported for the speech recognition components of a spoken dialog systems (Chapters 6, 7 and 34), 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.

Speech-to-text systems historically produce a case insensitive, unpunctuated output. Recently there have been a number of efforts to produce STT outputs with correct case and punctuation, as well as conversion of numbers, dates, and acronymns to a standard written form. This is essentially the reverse process of the text normalization steps described in Section 4. Both linguistic and acoustic information (essentially pause and breath noise cues) are used to add punctuation marks in the speech recognizer output. An efficient method is to rescore word lattices that have been expanded to permit punctuation marks after each word, sentences boundaries at each pause, with a specialized case sensitive, punctuated language model.

7 Discussion and Perspectives

Despite the numerous advances made over the last decade, speech recognition is far from a solved problem. Current reseach topics aim to develop generic recognition models and to use unannotated data for training purposes, in an aim to reduce the reliance on manually annotated training corpora.

Much of the progress in LVCSR has been fostered by supporting infrastructure for data collection, annotation and evaluation. The Speech Group at the National Institute of Standards and Technology (NIST) has been organizing benchmark evaluations for a range of human language technologies (speech recognition, speaker and language recognition, spoken document retrieval, topic detection and tracking, automatic content extraction, spoken term detection) for over 20 years, recently extended to also include related multi-modal technologies [4]

While the performance of speech recognition technology has dramatically improved for a number of 'dominant' languages (English, Mandarin, Arabic, French, Spanish, ...), generally speaking technologies for language and speech processing are available only for a small proportion of the world's languages. By several estimations there are over 6000 spoken languages in the world, but only about 15% of them are also written. Text corpora, which can be useful for training the language models used by speech recognizers, are becoming more and more readily available on the Internet. The site http://www.omniglot.com lists about 800 languages that have a written form.

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.

Todays systems often also provide additional information which is useful for structuring audio data. In addition to the linguistic message, the speech signal encodes information about the characteristics of the speaker, the acoustic environment, the recording conditions and the transmission

[4] See http://www.nist.gov/speech/tests.

channel. Acoustic metadata can be extracted from the audio to provide a description including the language(s) spoken, the speaker(s), accent(s), acoustic background conditions, the speaker's emotional state etc. Such information can be used to improve speech recognition performance, and to provide an enriched text output for downstream processing. The automatic transcription can also be used to provide information about the linguistic content of the data (topic, named entities, speech style, ...). By associating each word and sentence with a specific audio segment, an automatic transcription can allow access to any arbitrary portion of an audio document. If combined with other meta-data (language, speaker, entities, topics) access via other attributes can be facilitated.

A wide range of potential applications can be envisioned based on automatic annotation 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), and question-answering. 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. Another challenging problem is recognizing spontaneous speech data collected with far-field microphones (such as meetings and interviews), which have difficult acoustic conditions (reverberation, background noise) and often have overlapping speech from different speakers.

Further Reading and Relevant Resources

An excellent reference is "Corpus Based Methods in Language and Speech Processing," edited by Young and Bloothooft (1997). This book provides an overview of currently used statistically based techniques, their basic principles and problems. A theoretical presentation of the fundamentals of the subject is given in the book "Statistical Methods for Speech Recognition" by Jelinek (1994). A general introductory tutorial on HMMs can be found in Rabiner and Juang (1986). "Pattern Recognition in Speech and Language Processing"by Chou and Juang (2003) and and "Multilingual Speech Processing" by Schultz and Kirchhoff (2006), provide more advanced reading. For general speech processing reference, the classical book Digital Processing of Speech Signals (Rabiner and Shafer, Prentice Hall, 1978) remains

relevant. A comprehensive discussion on signal representation can be found in Chapter 1.3 of the Survey of the State of the Art in Human Language Technology (http://www.cslu.ogi.edu/HLTsurvey) The most recent work in speech recognition can be found in the proceedings of major conferences (IEEE ICASSP, ISCA Interspeech) and workshops (most notably DARPA, ISCA ITRWs, IEEE ASRU), as well as the journals on *Speech Communication* and *Computer Speech and Language*. In the latter journal a special issue in October 1998 was devoted to Evaluation in Language and Speech technology.

Several web sites of interest are:

European Language Resources Association (ELRA) http://www.icp.inpg.fr/ELRA/home.html

European Speech Communication Association (ESCA) http://www.esca-speech.org

Linguistic Data Consortium (LDC) http://www.ldc.upenn.edu/

NIST Spoken Natural Language Processing http://www.itl.nist.gov/div894/894.01

Survey of the State of the Art in Human Language Technology
http://www.cslu.ogi.edu/HLTsurvey

Languages of the world http://www.omniglot.com

OLAC: Open Language Archives Community http://www.language-archives.org

References

- Special issue on "Modeling pronunciation variation for automatic speech recognition," *Speech Communication*, H. Bourlard, S. Furui, N. Morgan, H. Strik, eds. **29**(2-4), Nov. 1999

- Special issue on "Pronunciation Modeling and Lexicon Adaptation," E. Fosler-Lussier, W. Byrne, D. Jurafsky, eds. *Speech Communication*, **46**(2), June 2005.

- Bahl, L.R., J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer, and H.F. Silverman. 1976. "Preliminary results on the performance of a system for the automatic recognition of continuous speech". *Proceedings of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP-76)*, **-**. Philadelphia.

- Bahl, L.R., P. Brown, P. de Souza, R.L. Mercer, and M. Picheny. 1988. "Acoustic Markov Models used in the Tangora Speech Recogni-

tion System". *Proceedings of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP-88).* **1**. 497-500. New York.

- 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.

- Chen, L., J.L. Gauvain, L. Lamel, and G. Adda, Dynamic Language Modeling for Broadcast News, *ICSLP'04*. pp. 1281-1284, Jeju Island, 2004.

- Cohen, M. 1989. *Phonological Structures for Speech Recognition.* PhD Thesis, University of California, Berkeley, U.S.A.

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

- 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.

- Dreyfus-Graf, J. 1949. "Sonograph and Sound Mechanics". *Journal of the Acoustic Society of America*, **22**. 731.

- Dudley, H. and S. Balashek. 1958. "Automatic Recognition of Phonetic Patterns in Speech". *Journal of the Acoustic Society of America.* **30**. 721.

- Fousek P., L. Lamel and J.L. Gauvain. 2008. "On the use of MLP features for broadcast news transcription." *TSD'08*, volume 5246 of *Lecture Notes in Computer Science*, pages 303–310.

- Gales, M. and S. Young. 1995. "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech & Language*, **9**(4). 289-307. October.

- Gauvain, J.L. and C.H. Lee. 1984. "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains". *IEEE Transactions on Speech and Audio Processing.* **2**(2). 291-298.

- Hermansky, H. 1990. "Perceptual linear predictive (PLP) analysis of speech". *Journal of the Acoustic Society America.* **87**(4). 1738-1752.
- Hermansky, H. and S. Sharma. 1998. "TRAPs - classifiers of TempoRAl Patterns." *ICSLP'98*, Sydney, **3**:1003-1006.
- Hunt, M.J. 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)
- Jelinek, F. 1976. "Continuous Speech Recognition by Statistical Methods". *Proceedings of the IEEE.* **64**(4). 532-556. April.
- Jiang, H., "Confidence measures for speech recognition: A survey." 2004. *Speech Communication*, **45**():455-470.
- Juang, B.-H. 1985. "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT&T Technical Journal*, **64**(6). **-**
- Katz, S.M. 1987. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer". *IEEE Trans. Acoustics, Speech, and Signal Processing.* **ASSP-35**(3). 400-401.
- Leggetter, C.J. and P.C. Woodland. 1995. "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models". *Computer Speech and Language.* **9**. 171-185.
- Liporace, L. R. 1982. "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Transactions on Information Theory*, **IT-28**(5). 729-734.
- Liu, X., M. J. F. Gales, P.C. Woodland, Context Dependent Language Model Adaptation, *Interspeech'08.* pp. 837-840, Brisbane, 2008.
- Mangu, L., E. Brill and A. Stolcke. 2000. "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer Speech and Language*, **14**:373–400, 2000.
- Oshika, B.T., V.W. Zue, R.V. Weeks, H. Neu, and J. Aurbach. 1975. "The Role of Phonological Rules in Speech Understanding Research".

IEEE Transasctions on Acoustics, Speech, Signal Processing. **ASSP-23**. 104-112.

- Rabiner, L.R., and R.W. Schafer. 1978. *Digital processing of speech signals*, Englewood Cliffs; London: Prentice-Hall.

- Rabiner, L.R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proc. IEEE*, **77**(2). 257-286. February.

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

- Riley, M.D., and A. Ljojle. 1996. "Automatic Generation of Detailed Pronunciation Lexicons". in *Automatic Speech and Speaker Recognition*, Kluwer Academic Pubs, Ch. 20. 285-301.

- Schwartz, R., Y. Chow, S. Roucos, M. Krasner, and J. Makhoul. 1984. "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition". *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'84).* San Diego, U.S.A. **3**. 35.6.1-35.6.4.

- Stolcke, A., B. Chen, H. Franco, V.R.R. Gadde, M. Graciarena, M.Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. 2006. "Recent innovations in speech-to-text transcription at SRI-ICSI-UW". *Audio, Speech and Language Processing, IEEE Transactions on*, 14(5):1729–1744.

- Vintsyuk, T.K. 1968. "Speech discrimination by dynamic programming," *Kibnernetika.* **4**. 81.

- Q. Zhu, A. Stolcke, B. Chen, and N. Morgan. 2005. "Using MLP features in SRI's conversational speech recognition system." In *Interspeech'05*, Lisbon, Portugal. pages 2141–2144.

8 Glossary

Acoustic model: a model describing the probabilistic behavior of the encoding of the linguistic information in a speech signal
Acoustic parameterization: selection of acoustic features which are used to reduce model complexity without losing relevant linguistic information
Allophones: the collection of different realizations of a given phoneme, such as the aspirated /t/ in "type", the flapped /t/ in "butter", or final unreleased /t/ in "hot"
Back-off: mechanism for smoothing the estimates of the probabilities of rare events by relying on less specific models
Context-dependent model; a model which takes into account the neighboring phones
Frame or **Parameter vector:** set of acoustic parameters associated with a windowed portion of the signal
Language model: model used to estimate the probability of word sequences
Lexicon: list of words known by the recognizer, each word associated with one or more pronunciations
Phone: commonly used elementary units which generally correspond to phonemes, but may also correspond to allophones
Recording channel: means by which the audio signal is recorded (direct microphone, telephone, radio, etc
Speech recognition: transcription of the speech signal into a sequence of words

9 Authors

Lori Lamel

Lori Lamel is a senior CNRS researcher in the Spoken Language Processing group at LIMSI which she joined in October 1991. She received her PhD degree in EECS in May 1988 from the Massachusetts Institute of Technology. Her principal research activities are in speech recognition; acoustic-phonetic studies; lexical and phonological modeling; and conversational systems. She has been a prime contributor to the LIMSI participations in speech recognizer benchmark evaluations and developed the American English pronunciation lexicon. She has been involved in many European projects related to speech processing. She is a member of the Speech Communication Editorial Board, was a member of the Interspeech International Advisory Council, the IEEE James L. Flanagan Speech and Audio Processing Award Committee (2006-2009) and the EU-NSF Working Group for 'Spoken-Word Digital Audio Collections'. She has over 200 reviewed publications, and is co-recipient of the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal.

Jean-Luc Gauvain

Jean-Luc Gauvain is a senior researcher at the CNRS, where he is head of the Spoken Language Processing Group at LIMSI. He received a doctorate in Electronics from the University of Paris-Sud 11 in 1982, and joined the CNRS as a permanent researcher in 1983. His primary research centers on large vocabulary continuous speech recognition and audio indexing. His research interests also include conversational interfaces, speaker identification, language identification, and speech translation. He has participated in many speech-related projects both at the French National and European levels and has led the LIMSI participation in DARPA/NIST organized evaluations since 1992, most recently for the transcription of broadcast news data and of conversational speech. He has over 240 publications and received the 1996 IEEE SPS Best Paper Award in Speech Processing and the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal. He was co-editor-in-chief of the Speech Communication Journal from 2007 to 2009. Since April 2008, he is the Scientific Coordinator for the Quaero programme.

Index list:
acoustic features
acoustic parameters
acoustic model
acoustic model training
adaptation
allophone
Baum-Welch reestimation
Bayesian adaptation
Cepstrum
Cepstral mean removal
confidence score
confusion network decoding
consensus decoding
context-dependency
decoding
delta parameters
dictation
Expectation-Maximization algorithm
feature extraction
Fourier transform
Gaussian
Hamming
hidden Markov model (HMM)
language model (LM)
lexicon
lexical coverage
linear predictive coding (LPC)
Markov model
maximum a posteriori (MAP)
Mel
Mel Frequency Cepstral Coefficients (MFCC)
Maximum Likelihood Linear Regression (MLLR)
Multi layer perceptrons (MLP)
n-gram
out-of-vocabulary (OOV)

perplexity
phone
phoneme
Perceptual Linear Prediction (PLP)
pronunciation
recording channel
sentence error rate (SER)
speaker adaptive training (SAT)
speaking rate
speaking style
spectral representation
statistical models
TempoRAl Patterns
Viterbi
word error rate (WER)

List of acronyms

DARPA: Defense Advanced Research Projects Agency

DP: Dynamic programming

EM: Expectation-Maximization

HMM: hidden Markov model

LDC: Linguistics Data Consortium

LM: language model

LPC: Linear Predictive Coding

LVCSR: large vocabulary continuous speech recognition

MAP: maximum a posteriori

MFCC: Mel Frequency Cepstral Coefficients

ML: maximum likelihood

MLLR: maximum likelihood linear regression

MLP: multi layer perceptrons

NIST: National Institute of Standards and Technology

OOV: out-of-vocabulary

PDF: probability density function

PLP: Perceptual Linear Prediction

SAT: speaker adaptive training

TRAPs: TempoRAl Patterns