Some Issues in Speech Recognizer Portability

Lori Lamel

Spoken Language Processing Group,
LIMSI-CNRS, France
lamel@limsi.fr

Abstract

Speech recognition technology has greatly evolved over the last decade. However, one of the remaining challenges is reducing the development cost. Most recognition systems are tuned to a particular task and porting the system to a new task (or language) requires substantial investment of time and money, as well as human expertise. Todays state-of-the-art systems rely on the availability of large amounts of manually transcribed data for acoustic model training and large normalized text corpora for language model training. Obtaining such data is both time-consuming and expensive, requiring trained human annotators with substantial amounts of supervision. This paper addresses some of the main issues in porting a recognizer to another task or language, and highlights some some recent research activities aimed a reducing the porting cost and at developing generic core speech recognition technology.

1. Introduction

Speech recognition tasks can be categorized by several dimensions: the number of speakers known to the system, the vocabulary size, the speaking style, and the acoustic conditions. Concerning speakers, the most restrictive is when only one speaker can use the system and the speaker is required to enroll with the system in order to be recognized (speaker-dependent). The system may be able to recognize speech from several speakers, but still requires enrollment data (multiple speaker) or the system can recognize the speech from nominally any speaker without any training data (speaker-independent).

A decade ago the most common recognition tasks were either small vocabulary isolated word or phrases or speaker dependent dictation, whereas today speech recognizers are able to transcribe unrestricted continuous speech from broadcast data in multiple languages with acceptable performance. The increased capabilities of todays recognizers is in part due to the improved accuracy (and increased complexity) of the models, which are closely related to the availability of large spoken and text corpora for training, and the wide availability of faster and cheaper computational means which have enabled the development and implementation of better training and decoding algorithms. Despite the extent of progress over the recent years, recognition accuracy is still quite sensitive to the environmental conditions and speaking style: channel quality, speaker characteristics, and background noise have a large impact on the acoustic component of the speech recognizer, whereas the speaking style and discourse domain largely influence the linguistic component. In addition, most systems are both task and language dependent, and bringing up a system for a different task or language is costly and requires human expertise.

Only for small vocabulary, speaker-dependent isolated word or phrase speech recognizers, such as name dialing on mobile telephones, portability is not really an issue. With such devices, all of the names must be entered by the user according to the specific protocol - such systems typically use whole word patterns and do not care who the speaker or what the language is. For almost all more complex tasks, portability is a major concern. Some speech technology companies have been addressing the language localization problem for many years, and some research sites have also been investigating speech recognition in multiple languages (4; 13; 14; 21; 35; 37) as well as speech recognition using multi-lingual components (19; 33). Multi-lingual speech processing has been the subject of several special sessions at conferences and workshops (see for example, (1; 2; 3; 20)). The EC CORETEX project (http://coretex.itc.it) is investigating methods to improve basic speech recognition technology, including fast system development, as well as the development of systems with high genericity and adaptability. Fast system development refers to both language support, i.e., the capability of porting technology to different languages at a reasonable cost; and task portability, i.e. the capability to easily adapt a technology to a new task by exploiting limited amounts of domain-specific knowledge. Genericity and adaptability refer to the capacity of the technology to work properly on a wide range of tasks and to dynamically keep models up to date using contemporary data. The more robust the initial generic system is, the less there is a need for adaptation.

In the next section an overview of todays most widely used speech recognition technology is given. Following subsections address several approaches to reducing the cost of porting, such as improving model genericity, and reducing the need for annotated training data. An attempt is made to give an idea of the amount of data and effort required to port to a different language or task.

2. Speech Recognition Overview

Speech recognition is concerned with converting the speech waveform into a sequence of words. Today's most performant approaches are based on a statistical modelization of the speech signal (16; 31; 32; 38). The basic modeling techniques have been successfully applied to a number of languages and for a wide range of applications.

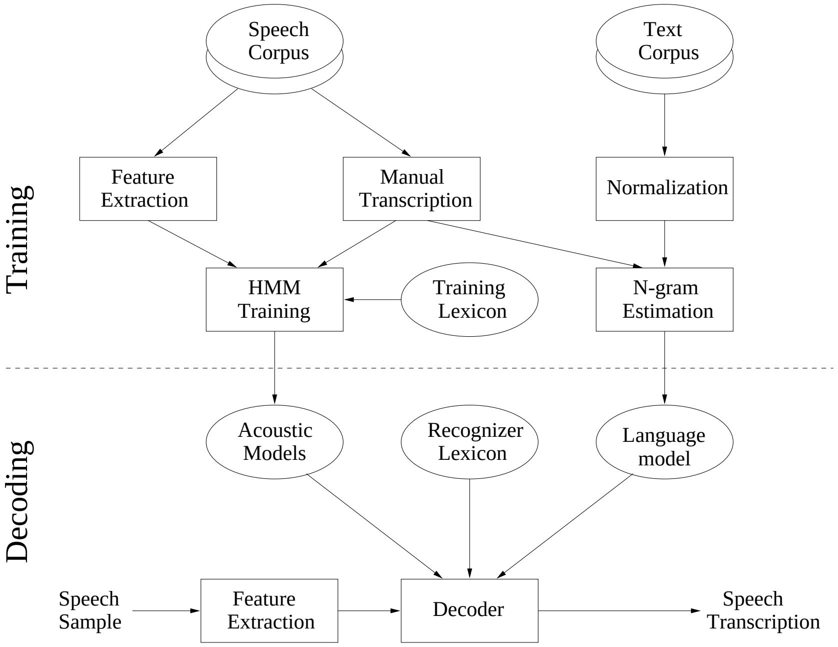

Figure 1: System diagram of a generic speech recognizer based using statistical models, including training and decoding processes.

The main components of a 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 training and decoding algorithms are largely task and language independent, the main language dependencies are in the knowledge sources (the training corpora).

The first step of the acoustic feature analysis is digitization, in which the continuous speech signal is converted into discrete samples. Acoustic feature extraction is then carried out on a windowed portion of speech [1], with the goal of reducing model complexity while trying to maintain the linguistic information relevant for speech recognition. 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. Cepstral mean removal (subtraction of the mean from all input frames) is commonly used to reduce the dependency on the acoustic recording conditions, and delta parameters (obtained by taking the first and second differences of the parameters in successive frames) are often used to capture the dynamic nature of the speech signal. While the details of the feature analysis differs from system to system, most of the commonly used analyses can be expected to work reasonably well for most languages and tasks.

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 (32). These models are popular as they are performant and their parameters can be efficiently estimated using well established techniques. The Markov model is described by the number of states and the transitions probabilities between states. The most widely used acoustic units in continuous speech recognition systems are phone-based[2], and typically have a small number of left-to-right states in order to capture the spectral change across time. Since the number of states imposes a minimal time duration for the unit, some configurations allow certain states to be skipped. The probability of an observation (i.e. a speech vector) is assumed to be dependent only on the state, which is known as the 1st order Markov assumption.

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” (a commonly used approach for isolated word recognition) or a different mapping such as the fenones (7). Compared with larger units, small subword units reduce the number of parameters, and more importantly can be associated with back-off mechanisms to model rare or unseen, contexts, and facilitate porting to new vocabularies. Fenones offer the additional advantage of automatic training which is of interest for language porting, 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 mono-

[1] An inherent assumption is that 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).

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

phone 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 or may not be merged. Different approaches can be used to select the contextual units based on frequency or using clustering techniques, or decision trees, and different types of contexts have been investigated. 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 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). 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. Speaker-independence is obtained by estimating the parameters of the acoustic models on large speech corpora containing data from a large speaker population. Since there are substantial differences in speech from male and female talkers arising from anatomical differences it is thus common practice to use separate models for male and female speech in order to improve recognition performance (requiring automatic gender identification).

2.1. Lexical and pronunciation modeling

The lexicon is the link between the acoustic-level representation and the word sequence output by the speech recognizer (34). 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. Developing a consistent pronunciation lexicon requires substantial language specific knowledge from a native speaker of the language and usually entails manual modification even if grapheme-to-phoneme rules are reasonably good for the language of interest. The lexical units must be able to be automatically extracted from a text corpus or from speech transcriptions and for a given size lexicon should optimize the lexical coverage for the language and the application. 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. The recognition word list is to some extent dependent on the conventions used in the source text (punctuation markers, compound words, acronyms, case sensitivity, ...) and the specific language. The lexical units can be chosen to explicitly model observed pronunciation variants, for example, using compound words to represent word sequences subject to severe reductions such as "dunno" for "don't know". The vocabulary is usually comprised of a simple list of lexical items as observed in the text. Attempts have been made to use other units, for example, to use a list of root forms (stems) augmented by derivation, declension, composition rules. However, while more powerful in terms of language coverage, such representations are more difficult to integrate in present state-of-the-art recognizer technology.

These pronunciations may be taken from existing pronunciation dictionaries, created manually or generated by an automatic grapheme-phoneme conversion software. Alternate pronunciations are sometimes used to explicitly represent variants that cannot be easily modeled by the acoustic units, as is the case for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as *excuse, record, produce*. While pronunciation modeling is widely acknowledged to be a challenge to the research community, there is a lack of agreement as to what pronunciation variants should be modeled and how to do so. Adding a large number of pronunciation variants to a recognition lexicon without accounting for their frequency of occurrence can reduce the system performance. An automatic alignment system is able to serve as an analysis tool which can be used to quantify the occurrence of events in large speech corpora and to investigate their dependence on lexical frequency (5).

2.2. 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 (17).

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.

One of the main motivations for text normalization is to reduce lexical variability so as to increase the coverage for a fixed vocabulary size. The 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*. However with increased interest in going beyond transcription to information extraction tasks (such as finding named entities or locating events in the audio signal) such distinctions are important. In our work at LIMSI for other languages (French, German, Portuguese) capitalization of proper names is distinctive with different lexical items for the French words *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*. Generally speaking, 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.

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. 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, where the interpolation weights are estimated on some development data.

2.3. Decoding

The aim of the decoder is to determine the word sequence with the highest likelihood given the lexicon and the acoustic and language models. Since it is often prohibitive to exhaustively search for the best solution, techniques have been developed to reduce the computational load by limiting the search to a small part of the search space. 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. Multi-pass decoding strategies progressively add knowledge sources in the decoding process and allows the complexity of the individual decoding passes to be reduced. Information between passes is usually transmitted via word graphs, although some systems use N-best hypotheses (a list of the most likely word sequences with their respectives scores). One important advantage of multi-pass is the possibility to adapt the models between decoding passes. 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. Attempts at language model adaptation have been less successful. However, multi-pass approaches are not well suited to real-time applications since no hypothesis can be returned until the entire utterance has been processed.

3. Language porting

Porting a recognizer to another language necessitates modification of some of the system parameters, i.e. those incorporating language-dependent knowledge sources such as the phone set, the recognition lexicon (alternate word pronunciations), and phonological rules and the language model. Different languages have different sets of units and different coarticulation influences amomg adjacent phonemes. This influences the way of choosing context-dependent models and of tying distributions. 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.

One important aspect in developing a transcription system for a different language is obtaining the necessary resources for training the acoustic and language models, and a pronunciation lexicon. The Linguistic Data Consortium (LDC http://www.ldc.upenn.edu) and the European Language Resources Association (ELRA http://www.elda.fr) have greatly aided the creation and distribution of language resources. The number and diversity of language resources has grown substantially over recent years. However, most of the resources are only available for the most interesting languages from the commercial or military perspectives.

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 (33). This approach offers the advantage of being able to use the 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.

A general rule of thumb for the necessary resources for speaker independent, large vocabulary continuous speech recognizers is that the minimal data requirements are on the order of 10 hours transcribed audio data for training the acoustic models and several million words of texts (transcriptions of audio if available) for language modeling. Depending upon the application, these resources are more or less difficult to obtain. For example, unannotated data for broadcast news type tasks can be easily recorded via standard TV, satellite or cable and data of this type is becoming more easily accessible via the Internet. Related text materials are also available from a variety of on-line newspapers and new feeds. The manual effort required to transcribe broadcast news data is roughly 20-40 hours per hour of audio data, depending upon the desired precision (8).

Data for other applications can be much more difficult to obtain. In general, for spoken language dialog systems, training data needs to be obtained from users interacting with the system. Often times an initial corpus is recorded from a human-human service (should it exist) or using simulations (Wizard-of-OZ) or an initial prototype system. The different means offer different advantages. For example, WOz simulations help in making design decisions before the technology is implemented and allow alternative designs to be simulated quickly. However, the amount of data that can be collected with a WOz setup is limited by the need for a human wizard. Prototype systems offer the possibility of collection much larger corpora, albeit somewhat limited by the capacity of the current system. We have observed that the system's response generation has a large influence on the naturalness of the data collected with a prototype system.

Other application areas of growing interest are the transcription of conversational speech from telephone conversations and meetings, as well as voicemail. Several sources of multilingual corpora are available (for example, the CallHome and CallFriend corpora from LDC). This data is quite difficult to obtain and costly to annotate due to its very spontaneous nature (hesitations, interruptions, use of jargon). The manual effort involved is higher than that required for broadcast news transcription, and the transcriptions are less consistent and accurate.

The application-specific data is useful for accurate modeling at different levels (acoustic, lexical, syntactic and semantic). Acquiring sufficient amounts of text training data is more challenging than obtaining acoustic data. With 10k queries relatively robust acoustic models can be trained, but these queries contain only on the order of 100k words, which probably yield an incomplete coverage of the task (ie. they are not sufficient for word list development) and are insufficient for training *n*-gram language models.

At LIMSI broadcast news transcription systems have been developed for the American English, French, German, Mandarin, Spanish, Arabic and Portuguese languages. The Mandarin language was chosen because it is quite different from the other languages (tone and syllable-based), and Mandarin resources are available via the LDC as well as reference performance results from DARPA benchmark tests. To give an idea of the resources used in developing these systems, the training material are shown in Table 1. It can be seen that there is a wide disparity in the available language resources for a broadcast news transcription task: for American English, 200 hours of manually transcribed acoustic training were available from the LDC, compared with only about 20-50 hours for the other languages. Obtaining appropriate language model training data is even more difficult. While newspaper and newswire texts are becoming widely available in many languages, these texts are quite different than transcriptions of spoken language. Over 10k hours of commercial transcripts are available for American English (from PSMedia), and many TV stations provide closed captions. Such data are not available for most other languages, and in some countries it is illegal to sell transcripts. Not shown here, manually annotated broadcast news corpora are also available for the Italian (30 hours) and Czech (30 hours) languages via ELRA and LDC respectively, and some text sources can be found on the Internet.

Some of the system characteristics are shown in Table 2, along with indicative recognition performance rates for these languages. State-of-the-art systems can transcribe unrestricted American English broadcast news data with word error rates under 20%. Our transcription systems for French and German have comparable error rates for news broadcasts (6). The character error rate for Mandarin is also about 20% (10). 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.

4. Reducing the porting cost

4.1. 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 (24; 9; 15; 26; 30; 11). In (26) broadcast news (BN) (28) acoustic and language models to decode the test data for three other tasks (TI-digits (27), ATIS (12) and WSJ (29)). 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 ob-

	Audio			*Text (words)*	
Language	*Radio-TV sources*	*Duration*	*Size*	*News*	*Com.Trans.*
English	ABC, CNN, CSPAN, NPR, PRI, VOA	200h	1.9M	790M	240M
French	Arte, TF1, A2, France-Info, France-Inter	50h	0.8M	300M	20M
German	Arte	20h	0.2M	260M	-
Mandarin	VOA, CCTV, KAZN	20h	0.7M(c)	200M(c)	-
Portuguese	9 sources	3.5h	$\sim$35k	70M	-
Spanish	Televisa, Univision, VOA	30h	0.33M	295M	-
Arabic	tv: Aljazeera, Syria; radio: Orient, Elsharq, ...	50h	0.32M	200M	-

Table 1: Approximate sizes of the transcribed audio data and text corpora used for estimating acoustic and language models. For the text data, newspaper texts (News) and commercial transcriptions (Com.Trans.) are distinguished in terms of the millions of words (or characters for Mandarin). The American English, Spanish and Mandarin data are distributed by the LDC. The German data come from the EC OLIVE project and the French data partially from OLIVE and from the DGA. The Portuguese data are part of the 5h, 11 source Pilot corpus used in the EC ALERT project (data from 2 sources 24Horas and JornalTarde were reserved for the test set). The Arabic data were produced by the Vecsys company in collaboration with the DGA.

		Lexicon		*Language Model*		*Test*	
Language	*#phon.*	*size (words)*	*coverage*	*N-gram*	*ppx*	*Duration*	*%Werr*
English	48	65k	99.4%	11M fg, 14M tg, 7M bg	140	3.0h	20
French	37	65k	98.8%	10M fg, 13M tg, 14M bg	98	3.0h	23
German	51	65k	96.5%	10M fg, 14M tg, 8M bg	213	2.0h	25(n)-35(d)
Mandarin	39	40k+5k(c)	99.7%	19M fg, 11M tg, 3M bg	190	1.5h	20
Spanish	27	65k	94.3%	8M fg, 7M tg, 2M bg	159	1.0h	20
Portuguese	39	65k	94.0%	9M tg, 3M bg	154	1.5h	40
Arabic	40	60k	90.5%	11M tg, 6M bg	160	5.7h	20

Table 2: Some language characteristics. Specified for each language are: 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 language model size and the perplexity, the test data duration (in hours) 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. For German, separate word error rates are given for broadcast news (n) and documentaries (d).

tain 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 (25).

In (9) 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 (30). 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 (11).

4.2. Reducing the need for annotated training data

With today's technology, the adaptation of a recognition system to a new task or new language requires the availability of sufficient amount of transcribed training data. When changing to new domains, usually no exact transcriptions of acoustic data are available, and the generation of such transcribed data is an expensive process in terms of manpower and time. On the other hand, there often exist incomplete information such as approximate transcriptions, summaries or at least key words, which can be used to provide supervision in what can be referred to as "informed speech

Amount of training data		*Language Model*
Raw	*Usable*	News.Com.Cap
10min	10min	53.1
1.5h	1h	33.3
50h	33h	20.7
104h	67h	19.1
200h	123h	18.0

Table 3: Supervised acoustic model training: Word error rate (%) on the 1999 evaluation test data for various conditions using one set of gender-independent acoustic models trained on subsets of the HUB4 training data with detailed manual transcriptions. The language model is trained on the available text sources, without any detailed transcriptions of the acoustic training data. The raw data reflects the size of the audio data before partitioning, and the usable data the amount of data used in training the acoustic models.

	Raw Acoustic training data		
Language model	*200 hours*	*1.5 hours*	*10 min*
News.Com.Cap, 65k	18.0	33.3	53.1
News, 65k	20.9	36.1	55.6
30 M words, 60k	24.1	40.8	60.2
1.8 M words, 40k	28.8	46.9	65.3

Table 4: Supervised acoustic model training: Reference word error rates (%) on the 1999 evaluation test data with varying amounts of manually annotated acoustic training data and a language model trained on 1.8 M and 30 M words of news texts from 1997.

Raw Acoustic training data		*WER (%)*
bootstrap models	10 min manual	65.3
1 (6 shows)	4 h	54.1
2 (+12 shows)	12 h	47.7
3 (+23 shows)	27 h	43.7
4 (+44 shows)	53 h	41.4
5 (+60 shows)	103 h	39.2
6 (+58 shows)	135 h	37.4

Table 5: Unsupervised acoustic model training: Word error rate (%) on the 1999 evaluation test data with varying amounts of automatically transcribed acoustic training data and a language model trained on 1.8 M words of news texts from 1997.

recognition”. Depending on the level of completeness, this information can be used to develop confidence measures with adapted or trigger language models or by approximate alignments to automatic transcriptions. Another 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 data base for the new domain can be cumulatively extended over time *without* direct manual transcription. This approach has been investigated in (18; 22; 23; 36; 39).

In order to give an idea of the influence of the amount of training data on system performance, Table 3 shows the performance of a 10xRealTime American English BN system for different amounts of manually annotated training data. The language model News.Com.Cap is trained on large text corpora, and results from the interpolation of individual language models trained on newspaper and newswires tests (790M words), commercially produced transcripts and closed-captions predating the test epoch (240M words). The word error is seen to rapidly decrease initially, with only a relatively small improvement above 30 hours of usable data. However, there is substantial information available in the language models. Table 4 summarizes supervised training results using substantially less language model training material. The second entry is for a language model estimated only on the newpaper texts (790M words), whereas for the remaining two language models were estimated on only 30 M words of texts (the last 2 months of 1997) and 1.8 M words (texts from December 26-31, 1997). It can be seen that the language model training texts have a large influence on the system performance, and even 30 M words is relatively small for the broadcast news transcription task.

The basic idea of light supervision is to use a speech recognizer to automatically transcribe unannotated data, thus generating “approximate” labeled training data. By iteratively increasing the amount of training data, more accurate acoustic models are obtained, which can then be used to transcribe another set of unannotated data. The manual work is considerably reduced, not only in generating the annotated corpus but also during the training procedure, since it is no longer necessary to extend the pronunciation lexicon to cover all words and word fragments occurring in the training data. In (22) it was found that somewhat comparable acoustic models could be estimated on 400 hours automatically annotated data from the TDT-2 corpus and 150 hours of carefully annotated data.

The effects of reducing the amount of supervision are summarized in Table 5. The first observation that can be made, is that even using a recognizer with an initial word error of 65% the procedure is converging properly by training acoustic models on automatically labeled data. This is even more surprising since the only supervision is via a language model trained on a small amount of text data predating the raw acoustic audio data. As the amount of automatically transcribed acoustic data is successively doubled, there are consistent reductions in the word error rate. While these error rates are still quite high compared to supervised training, retranscribing the same data (36) can be expected to reduce the word error rate further. (Recall that even with supervised acoustic model training trained on 200 hours of raw data the word error rate is 28.8% with this language model.)

4.3. Unsupervised Cross-Task Adaptation

An incremental unsupervised adaptation scheme was investigated for cross-task adaptation from the broadcast news task to the ATIS task (26). In this system-in-loop adaptation scheme, a first subset of the training data is automatically transcribed using the generic system. The acoustic and linguistic models of the generic system are then adapted with these automatically annotated data and the resulting models are used to transcribe another portion of the training data. One obvious use of this scheme is for online model adaptation in a dialog system.

Using about one-third (15 hours) of the ATIS training corpus transcribed with a BN system to adapt both the acoustic and language models, the word error rate is reduced from 20.8% to 6.9%. Transcribing the remaining data, and readapting the models reduces the word error to 5.5% (which can be compared to 4.7% for a task-specific system). Contrastive experiments have shown that this gain is somewhat equally split between adaptation of the acoustic and language models.

4.4. Cross Language Portability

The same basic idea was used to develop BN acoustic models for the Portuguese language for which substantially less manually transcribed data are available. RTP and INESC, partners in the Alert project (http:alert.uni-duisburg.de) provided 5 hours of manually annotated data from 11 different news programs. Two of the programs (82 minutes) were reserved for testing purposes (JornalTarde_20_04_00 and 24Horas_19_07_00). The remaining 3.5 hours of data were used for acoustic model training. The language model texts were obtained from the following sources: the Portuguese Newswire Text Corpus distributed by LDC (23M words from 1994-1998); Correio da Manha (1.6M words), Expresso (1.9M words from 2000-2001), and Jornal de Noticias (46M words, from 1996-2001), The recognition lexicon contains 64488 words. The pronunciations are generated by grapheme-to-phoneme rules, and use 39 phones.

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.

5. Conclusions

This paper has discussed the main issues in speech recognizer development and portability across languages and tasks. Today's most performant systems make use of statistical models, and therefore require large corpora for acoustic and language model training. However, acquiring these resources is both time-consuming, costly, and may be beyond the economic interest for many languages. Research is underway to reduce the need for manually annotated training data, thus reducing the human investment needed for system development when porting to another task or language. By eliminating the need for manual transcription, automated training can be applied to essentially unlimited quantities of task-specific training data.

The pronunciation lexicon still requires substantial manual effort for languages without straightfoward letter-to-sound correspondences, and to handle foreign words and proper names. For languages or dialects without a written form, the challenge is even greater, since important language modeling data are simply unavailable. Even if a transliterated form can be used, it is likely to be impractical to transcribe sufficient quantities of data for language model training.

In summary, our experience is that although general technologies and development strategies appear to port from one language to another, to obtain optimal performance language specificities must be taken into account. Efforts underway to improve the genericity of speech recognizers, and to reduce training costs will certainly help to enable the development of language technologies for minority languages and less economically promising applications.

Acknowledgments

This work has been partially financed by the European Commission and the French Ministry of Defense. I would like to thank the members of the Spoken Language Processing group at LIMSI who have participated to the research reported here and from whom I have borrowed experimental results.

REFERENCES

[1] *IEEE Workshop on Automatic Speech Recognition*, Special session on Multilingual Speech Recognition, Snowbird, Dec. 1995.

[2] *ICSLP'96*, Special session on "Multilingual Speech Processing," Philadelphia, PA, Oct. 1996.

[3] *Multi-Lingual Interoperabilty in Speech Technology*, RTO-NATO and ESCA ETRW, Leusden, Holland, Sept. 1999.

[4] M. Adda-Decker, "Towards Multilingual Inoperability in Speech Recognition," *Multi-Lingual Interoperabilty in Speech Technology*, RTO-NATO and ESCA ETRW, Leusden, Holland, 69-76, Sept. 1999.

[5] M. Adda-Decker, L. Lamel, "Pronunciation Variants Across Systems, Languages and Speaking Style," *Speech Communication*, "Special Issue on Pronunciation Variation Modeling", **29**(2-4): 83-98, Nov. 1999.

[6] M. Adda-Decker, G. Adda, L. Lamel, "Investigating text normalization and pronunciation variants for German broadcast transcription," *ICSLP'2000*, Beijing, China, Oct. 2000.

[7] L.R. Bahl, P. Brown, P. de Souza, R.L. Mercer, M. Picheny, "Acoustic Markov Models used in the Tangora Speech Recognition System," *ICASSP-88* **1**, pp. 497-500.

[8] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1-2): 5-22, Jan. 2001.

[9] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, D. Giuliani, "From Broadcast News to Spontaneous Dialogue Transcription: Portability Issues," *ICASSP'01*, Salt Lake City, May 2001.

[10] L. Chen, L. Lamel, G. Adda, J.L. Gauvain, "Broadcast News Transcription in Mandarin," *ICSLP'2000*, Beijing, China, Oct. 2000.

[11] R. Cordoba, P. Woodland, M. Gales "Improved Cross-Task Recognition Using MMIE Training" *ICASSP'02*, Orlando, Fl, May 2002.

[12] D. Dahl, M. Bates *et al.*, "Expanding the Scope of the ATIS Task : The ATIS-3 Corpus," *ARPA Spoken Language Systems Technology Workshop*, Plainsboro, NJ, 3-8, 1994.

[13] C. Dugast, X. Aubert, R. Kneser, "The Philips Large Vocabulary Recognition System for American English, French, and German," *Eurospeech'95*, 197-200, Madrid, Sept. 1995.

[14] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, J. Sakai, S. Seneff, V. Zue, "Multilingual spoken language understanding in the MIT Voyager system," *Speech Communication*, **17**(1-2): 1-18, Aug. 1995.

[15] D. Giuliani, M. Federico, "Unsupervised Language and Acoustic Model Adaptation for Cross Domain Portability" *ISCA ITRW 2001 Adaptation Methods For Speech Recognition*, Sophia-Antipolis, France, Aug. 2001.

[16] F. Jelinek, "Statistical Methods for Speech Recognition," Cambirdge: MIT Press, 1997.

[17] 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.

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

[19] J. Köhler, "Language-adaptation of multilingual phone models for vocabulary independent speech recognition tasks," *ICASSP'98*, **I**, 417-420, Seattle, May 1998.

[20] J. Kunzmann, K. Choukri, E. Janke, A. Kiessling, K. Knill, L. Lamel, T. Schultz, S. Yamamoto, "Portability of ASR Technology to new Languages: multilinguality issues and speech/text resources," slides from the panel discussion at *IEEE ASRU'01*, Madonna di Campiglio, Dec. 2001. (http://www.cs.cmu.edu/t̃anja/Papers/asru2001.ppt)

[21] L. Lamel, M. Adda-Decker, J.L. Gauvain, G. Adda, Spoken Language Processing in a Multilingual Context," *ICSLP'96*, 2203-2206, Philadelphia, PA, Oct. 1996.

[22] L. Lamel, J.L. Gauvain, G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer, Speech & Language*, Jan. 2002.

[23] L. Lamel, J.L. Gauvain, G. Adda, "Unsupervised Acoustic Model Training," *IEEE ICASSP'02*, Orlando, Fl, May 2002.

[24] L. Lamel, F. Lefevre, J.L. Gauvain, G. Adda, "Portability issues for speech recognition technologies," *HLT'2001*, 9-16, San Diego, March 2001.

[25] F. Lefevre, J.L. Gauvain, L. Lamel, "Improving Genericity for Task-Independent Speech Recognition," *EuroSpeech'01*, Aalborg, Sep. 2001.

[26] F. Lefevre, J.L. Gauvain, L. Lamel, "Genericity and Adaptability Issues for Task-Independent Speech Recognition," *ISCA ITRW 2001 Adaptation Methods For Speech Recognition*, Sophia-Antipolis, France, Aug. 2001.

[27] R.G. Leonard, "A Database for speaker-independent digit recognition," *ICASSP*, 1984.

[28] D.S. Pallett, J.G. Fiscus, *et al.* "1998 Broadcast News Benchmark Test Results," *DARPA Broadcast News Workshop*, 5-12, Herndon, VA, Feb. 1999.

[29] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP'92*, Kobe, Nov. 1992.

[30] D. Povey, P. Woodland, "Improved Discriminative Training Techniques For Large Vocabulary Continuous Speech Recognition", *IEEE ICASSP'01*, Salt Lake City, May 2001.

[31] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**(2): 257-286. Feb, 1989.

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

[33] T. Schultz, A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, **35** (1-2): 31-51, Aug. 2001.

[34] F. Van Eynde, D. Gibbon, eds., *Lexicon Development for Speech and Language Processing*, Dordrecht: Kluwer, 2000.

[35] A. Waibel, P. Geutner, L. Mayfield Tomokiyo, T. Schultz, M. Woszczyna, "Multilinguality in Speech and Spoken Language Systems," *Proceedings of the IEEE*, Special issue on Spoken Language Processing, **88**(8): 1297-1313, Aug. 2000.

[36] F. Wessel, H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *ASRU'01*, Madonna di Campiglio, Italy, Dec. 2001.

[37] S. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. van Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual Large Vocabulary Speech Recognition: The European SQALE Project," *Computer Speech and Language*, **11**(1): 73-89, Jan. 1997.

[38] S. Young, G. Bloothooft, eds., "Corpus Based Methods in Language and Speech Processing," Dordrecht: Kluwer, 1997.

[39] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, 301-305, Feb. 1998.