Sub-word-based language models for speech recognition: implications for spoken document retrieval

Martha Larson
GMD German National Research Center for Information Technology
Institute for Media Communication
and Dept. of Computer Science, Faculty of Electrical Engineering, Duisburg University
larson@gmd.de

Large Vocabulary Continuous Speech Recognition (LVCSR) is dependent on language models to constrain the acoustic search space by delivering an a priori probability of possible word sequences. A language model for LVCSR models a spoken document as a time series; it predicts language as a sequence of units drawn from a fixed alphabet.

The classic LVCSR language model is an n-gram model that models a language stream with the probabilities of words and sequences of contiguous words. The probability of each word is predicted by its n-1 predecessors. This probability is estimated on text training data using maximum likelihood. Back-off to lower order n-grams, or interpolation with lower order models provides probability estimates for events unseen in the training data.

Two divergences from garden-variety n-gram language models are popular and gaining ground in LVCSR, the integration of global constraints and the replacement of the traditional orthographic word base unit with sub-word base units. These two approaches enhance system performance and improve the capacity of the speech recognition system to handle out of vocabulary (OOV) words, words it has not been explicitly designed to recognize.

Language models for information retrieval (IR) have a different purpose than language models for LVCSR, namely they model the information content of a document. Language models for IR can often safely discard aspects of language related to form, like word sequence or exact morphology. For IR purposes a document might be adequately modeled with a vector of indexing features that is a histogram of word stems.

A Spoken Document Retrieval (SDR) system requires language representations suited for speech recognition as well as for IR. For additional general information concerning SDR refer to [4]. The first step in the design of any language model involves the choice of what units to use as fundamental features. For LVCSR language models these features are the underlying inventory of base units over which the statistical model is defined. For IR language models these features are the indexing features that will be used to compute the relevance of the document to the user query. This extended abstract explores the base units of language models as a point of contact between language models for LVCSR and for IR, and tries to shed light how SDR systems can be built with an optimal interface between speech recognition and IR.

The first section sketches distinction between acoustic units (typically phonemes) and language model units (typically words) in a classical speech recognition system. In the second section I discuss why sub-word units are helpful for LVCSR. I go on in the third section to survey SDR systems built on recognition systems using different base units (sub-phoneme units, phoneme-level units, phoneme-string units, syllable-level units, morpheme-level units), indicating relevant references and comment on their performance or potential. The fourth and final section discusses issues that should be addressed in the design of a language model unit inventory for SDR, with special reference to the case of German.

Acoustic base units vs. language model base units

Spoken Document Retrieval is a case of retrieval on corrupt data. The nature of the noise present in the string output by the speech recognizer is a function of the kind of underlying acoustic unit the recognition system is based on and what kind of language model it uses to constrain possible sequences of this underlying acoustic unit.

These two distinct levels of constraint mean that there are two distinct ways in which a LVCSR system can be considered to be sub-word based. The basic inventory of acoustic units can be sub-word units and/or the basic inventory of units of the language model can be sub-word units. Systems make this distinction between acoustic units and language model units, since the acoustic models and language models are trained separately. Acoustic models are trained on a typically smaller (in the order of 60 hours) set of tagged acoustic data, where language models are trained on a larger set of text data (up to 600M word tokens would be typical). This segregation is dictated by data availability constraints.

I discuss the difference between acoustic units and language model units in detail here, since the output of a SDR system built on a recognizer which uses phonemes as its acoustic base unit, will be noisy only at the word level, unless the language model base unit is also the phone. As [3] points out, OCR systems differ from LVCSR systems in that most OCR systems are phoneme based and recognition errors result in non-words. The typical LVCSR system uses orthographic words as its language model base units. When presented with a new proper name that is not in the lexicon, the system outputs a probable word in its place. If the system is more sophisticated, it can hypothesize that the word is one it doesn't know, but can give no further details. The system can never output a string of phonemes

not in the dictionary. A LVCSR system can capture no articulation at the sub-word level, unless it uses sub-word units as the language model units.

Acoustic base units A typical LVCSR system uses phones or tri-phones as acoustic units. Systems that implement other acoustic units, syllables or mixed-units, are the ones driving the state-of-the-art forward in this area. The bottleneck is getting enough acoustic training data. The larger the sub-word unit, the larger the sub-word unit inventory and collecting enough acoustic tokens to train each sub-word unit as a separate acoustic model is difficult.

Significant headway with syllable-based systems is being made by ISIP as reported in [6]. The advantage of training syllable models is that pronunciation variation is trained right into the acoustic model, and does not need to be modeled separately in the dictionary. Syllable models also automatically capture co-articulation effects. ISIP overcomes the data problem (almost 90% if their syllable inventory has less than 100 tokens in the training data) by implementing a system with mixed acoustic units, training syllable models where they can, and otherwise using phoneme models to build syllables in the lexicon. If you were going to build a SDR system on this recognizer, however, it would be important to realize that the base unit of the language model is the syllable, and that the output will be characterized by noise on the syllable level.

Language mode base units The typical LVCSR system uses the orthographic word as the basic unit of the language model. Systems that deviate from this norm continue to apply the word "lexicon" or "dictionary" to designate the inventory of basic language model units, the alphabet over which the language model is trained. The lexicon forms the basis of the pronunciation dictionary, which specifies the representation of each of the language model units in terms of acoustic units. Innovative systems are those systems that eschew the orthographic word as the basic unit of the language model, and instead choose morphemes or other sub-word units created through data driven processes. Such systems must be carefully optimized to maintain the high degree of language model constraint necessary for good decoder performance.

It has also proven to be advantageous to recognition accuracy to combine orthographic words into super-word lexicon entries. Such entities would without doubt yield features useful to IR, but I won't discuss them further here, except to say that they can be reduced, if necessary, into words without any complex processing effort.

The appeal of the sub-word unit for LVCSR

The phoneme acquired its status as the preferred underlying unit of acoustic modeling in speech recognition through intuitive appeal and sheer convenience. An audio file can be ground truthed with phonemes using a text transcription and a pronunciation dictionary. Modeling that goes on below the phoneme level bootstraps from the phoneme transcription, but remains system internal and is not part of the recognizer output.

The success of the phoneme language model in the area of handwriting recognition [1] does not carry over to LVCSR. Phoneme language models are too unconstrained to prove useful for transcription tasks. A good phoneme system achieves an error rate of 25% and in the real world error rates of up to 50% are not uncommon.

The orthographic word as the base unit for language modeling is probably a result of the historical origin of n-gram language models in English language speech recognition research. It is a conveniently available unit, which conventional wisdom grants the status of a useful information bearing entity. But is the orthographic word the best building block to use to represent all languages?

English has relatively little inflectional morphology (endings expressing case number and gender agreement) and prefers to express complex concepts as a phrase, or a hyphenated compound, rather than as a closed compound. English is, however, not the ideal language to research on if results need to generalize to compounding and highly inflected languages. Any other of the 11 official languages in the European Union (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish) exhibits a greater degree of compounding/inflection than English.

In correction of what might have been an initial misstep in the field, language models for LVCSR have been steadily freeing themselves from the orthographic word as the basic unit. In order to train a language model on base units other than the word, you have to re-tokenize the training text in to the new sub-word units. If your decomposition process is dictionary based, it might have just as many problems with out of vocabulary (OOV) words as the original recognition system. But if you have a good decomposition method, you can generate a sub-word-based language model with virtually infinite coverage.

German is a particularly challenging case, being an inflecting as well as compounding language. A lexicon for German language speech recognition needs to contain 3-4 times as many forms as that for English language speech recognition in order to attain the same coverage. A French language lexicon achieves the same coverage as the English language lexicon with only a doubled size [27].

Sub-word language model units alleviate the problem of the bloated lexicon [13][20]. But as the base units of the language model become fewer and smaller, the language model becomes less constrained. Base units must be

optimized to achieve maximum coverage and minimum ill effects of decreased contextual information. A typical word-based LVCSR system achieves 65%-75% word accuracy and ideal sub-word units would not degrade this performance significantly.

Integrating LVCSR into IR

There are two basic different strategies to integrate speech recognition and information retrieval. The first strategy is to use the recognizer to produce a transcript as close as possible to what was said in the speech document [15]. SDR then reduces to text IR, except for the noise in the transcription produced by recognizer errors.

The advantage of this first method, is that the transcript is readable by humans and can be used to represent the hit list. The user can skim the list without having to listen to each selection; audio is only scanable in (near) real time. The transcript can also be stored compactly and the original recordings may no longer even be necessary. Representing retrieval results for spoken audio can be tricky. [23] and especially [8] address this issue.

The disadvantage of creating a transcription before doing retrieval, is that a recognizer processing speech in transcription mode aims to minimize word error rate (WER), and this is not necessarily good for IR. WER concerns only the best hypothesis that the recognizer generates, discarding information contained in lower ranking hypotheses [19]. In addition, WER treats all words equally, whereas some have higher value for IR [23].

The second strategy for integrating speech recognition and information retrieval is to have the recognizer output information useful for the retrieval process, but not necessarily meaningful to human users. Under this strategy I distinguish two categories, approaches that use keyword spotting, a recognition mode in which no transcripts are generated, and the task of the recognizer is to extract key elements only [23][7], and approaches that leverage the N-best list or other representation of the N most likely hypotheses generated by the recognizer [5][18][21][28].

The following sections describe various sub-unit types and sketch their implementation in SDR.

Sub-phoneme language models The lowest conceivable level to do IR is on the signal level. The feature vectors extracted from the signal are perhaps a more useful low-level information source. Document classification has been successfully implemented using quantized feature vectors [25]. Such and approach is not be uninteresting for SDR.

Phoneme-level language models The most popular sub-word unit to be applied to SDR is the phoneme. A simple approach implements a recognizer that weights every phoneme as equally probable (does not implement a n-gram language model) and performs retrieval on the resulting N-best lattices [10]. Other SDR systems are built on top of recognizers using n-gram phone-based language models [25].

A more sophisticated system is described in [2]. The LVCSR systems builds a phoneme lattice which is not searched directly for keywords, but rather used in a rough match to identify time locations that a keyword was likely. Where a location in the lattice is identified as a potential keyword, the corresponding speech signal is then subjected to a detailed acoustic match. Another approach is presented in [26], where a phoneme recognizer is used to localize “slots”, contexts in which key words are likely to occur.

Both [22] and [16] suggest that phoneme-based SDR cannot compete with SDR based on very large vocabulary word-based speech recognition. Phoneme output tends to be noisy. It is tricky to “stem” in phoneme strings, unless the base form is pronounced the same in all inflections [26]. Sub-word language models based on units larger than phonemes may offer an optimal middle ground.

Phoneme-string language models Phoneme strings are the sub-word unit next up in size from phones. [17][18][19] investigate the use of variable length phonetic sequences, both in the language model for speech recognition and as the basis of (English language) information retrieval. Overlapping fixed-length phoneme sequences from 1 to 5 phones in length prove especially effective. Longer sequences quickly lead to degraded retrieval, since they overfit the noisy recognizer output. [18] also tests retrieval using phoneme classes, syllables and multigrams (data-derived non-overlapping phoneme strings of varying lengths), which are all interesting alternatives. These latter techniques do not achieve the performance of the overlapping phonetic sequences.

Syllable-level language models Although [18] reports second-rate results for syllables, syllables should not be disgarded as potential indexing features. The results that syllable-level models yield are without doubt influenced by additional factors, such as domain, language and acoustic unit being used for decoding. In [7] a system for German language spoken document retrieval that uses syllable-like (VCV) acoustic units is described. The system extracts indexing features, also at the syllable level, by processing the spoken documents with a speech recognizer in keyword spotting mode. This strategy outperforms the word-based system used as the baseline.

Morpheme-level language models Language models that use a morpheme-sized or morpheme-like base unit have not been extensively tested in SDR. Chinese language information retrieval has applied character-based units [24], but I am not aware of a parallel trend for Indo-European languages. The advantage of a morpheme-based language model is that morphemes encode semantics in an intuitively appealing manner. Also the speech recognizer can output a string from which the string forms can be directly derived. Morpheme language models are not really popular for

speech recognition, most likely because inflectional morphology tends to be short and thus acoustically confusable. Morpheme based inventories are disadvantaged by that fact that their units range in size from monosyllabic endings to polysyllabic roots. The ability of the recognizer to correctly decode inflectional (functional) morphology is, however, probably not important for IR. A possible strategy to compensate for these factors would be to weight longer words more heavily [26]. Longer words are more important to retrieval and are recognized with a high probability since they provide more contextual constraint.

Choosing indexing features for compounding/inflectional languages: the case of German

A list of guidelines for developing ideal indexing features for SDR is developed in [7]. Two important specifications stand at the top of that list. Indexing features for SDR must do a good job of discriminating between documents and that they must be reliably recognizable by the recognizer. Both of these criteria are dependent on the domain and the language(s) of the document collection, as well as on the level of difficulty of the task. This section takes the case of German and discusses some of the issues that must be considered in choosing language model base units and determining that they are suitable both as indexing features and as base units of the LVCSR language model.

Let's look at some specifics of the German language. German is infamous for its compounding. The compound word *Wettervorhersage* means "weather forecast" (German nouns are capitalized by convention). If we split *Wettervorhersage* on the morpheme level we get *Wetter* "weather" *vorhersage* "forecast". Splitting at the morpheme level tends to diffuse the semantic connections expressed through juxtaposition, much in the same way it would in English. Arguably the decomposition of *automatisch* "automatic" into *auto* and *matisch* is of questionable utility.

The natural conclusion is that certain words should be decomposed and certain words should not. It has also been shown to be beneficial for LVCSR if language model basic units are not created by splitting across the board, but rather by introducing selective splitting criteria [13]. The snag is that the LVCSR benefits from compound decomposition only where necessary to improve language model vocabulary coverage. Recognition performance is degraded when compounds that occur frequently are split, or when compound splitting leads to units that are small and/or have a high degree of acoustic confusability. IR benefits when decomposition gives access to units that enhance similarity in related documents. It remains to be determined, if the same compound decomposition is helpful for both LVCSR and IR.

If we split on the syllable level we get the stream *wet er vor her sag e.* A list of the most obvious alternative interpretations possible for the resulting units, gives an impressionistic idea of how the semantics of the original word is diluted, *Wet* is the stem of the verb "to bet", *er*, means "he", *vor* is the preposition "before" and also common prefix, *her* is the adverb "towards" and common prefix and *sag* is the stem of the verb "to say". Stop word removal is tricky when the recognizer outputs such a string. A naïve stop word elimination algorithm would leave us with "bet say", a feature pair not obviously closely associated with either weather or prediction.

If we derive features at the phone-string level, we get something like *wet ete ter erv rvo rhe her ers rsa sag age*. Such fine-grained decomposition makes each individual feature relevant to more queries, but the hope is that taken as a whole, these features will still prove specific enough to discriminate documents. Such a feature string also does not lend itself to removing stop words or stemming in the traditional way. An alternative might be, to train a system of weights. Phoneme strings highly likely to have originated from stop words or functional morphology are given low weights to minimize their contribution to the calculation of the distance between the document and the user query.

Stop word removal and stemming might also turn out to be unnecessary. [9] and [12] suggest that the benefits of stemming for IR is highly language dependent. [14] concludes that when document classification is implemented with a Support Vector Machine, known for its ability to deal with high dimensional data, stemming and the removal of stop words is not necessary.

This discussion has presented an overview of the role that sub-word units play in speech recognition and how those same sub-word units might also be used for information retrieval in order to create an optimized SDR system. My main message is that sub-word language model base units must be chosen in such a way that spoken documents of different topics are maximally discriminated for the given language and task. It is important that the LVCSR language model and IR language model be optimized together.

I would like to close with a couple additional observations. Why is it important that the same sub-units output by the recognizer also be the features of the indexing? For one, because it is tricky to reconstitute smaller units in an output stream in to larger ones, since there is not always a unique way of doing so. The second reason is because using the same features for both LVCSR and IR assures that the integration of information provided by recognizer confusion statistics is straightforward. Confusion matrices encoding the probabilities of the recognizer making certain mistakes can be used to expand the documents or user queries. Text or query expansion has been used to try to overcome the noise introduced by recognizer error [22]. This technique is often used on the phoneme level [3] [26], and has been implemented on the word level as well [21]. Its usefulness for phoneme-string and morpheme-level sub-word units has yet to be tested. A promising approach to the integration of LVCSR and IR, and one that will potentially enhance

system resistance to recognizer noise, is the combination of multimodal information sources, such a system combines the output of LVCSR systems based on different sub-word units [11] [24].

References

[1] A. Brakensiek, D. Willett, and G. Rigoll, "Unlimited vocabulary script recognition using character n-grams," *2. DAGM-Symposium, Tagungsband*, Springer-Verlag, 2000.

[2] S. Dharanipragada and S. Roukos, "New Work Detection in Audio Indexing," *ASRU* 1997.

[3] A. Ferrieux and S. Peillon, "Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval," ESCA ETRW workshop *Accessing information in spoken audio*, Cambridge, April 1999.

[4] J. Foote, "An Overview of Audio Information Retrieval, "*ACM-Springer Multimedia Systems*, 1998.

[5] J.T. Foote, S.J.Young, G.J.F. Jones and K. Sparck Jones, "Unconstrained keyword spotting using phone lattices with application to spoken document retrieval," *Computer Speech and Language*, 1997.

[6] A, Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, May 2001.

[7] U. Glavitsch and P. Schäuble. "A system for retrieving speech documents," *SIGIR* 1992.

[8] J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira and A. Singhal, "Finding information in audio: A new paradigm for audio browsing/retrieval," ESCA ETRW workshop *Accessing information in spoken audio*, Cambridge, April 1999. http://svr-www.eng.cam.ac.uk/~ajr/esca99/

[9] C. Jacquemin and E. Tzoukermann, "NLP for term variant extraction: synergy between morphology, lexicon and syntax," In T. Strzalkowski, ed. *Natural Langue Information Retrieval* Kluwer, Dordrecht 1999.

[10] D.A. James and S.J. Young. "A fast lattice-based approach to vocabulary independent word spotting," *ICASSP* 1994.

[11] G. J. F. Jones, J.T. Foote, K. Sparck Jones and S.J. Young. "Retrieving Spoken Documents by Combining Multiple Index Sources," *SIGIR* 1996.

[12] W. Kraaij and R. Pohlmann, "Viewing stemming as recall enhancement," *SIGIR* 1996.

[13] M. Larson, D. Willett, J. Koehler, G. and Rigoll, "Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches," *ICSLP* 2000.

[14] E. Leopold and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?" paper accepted for publication in *Machine Learning*. http://ais.gmd.de/KD/textmining.html

[15] J. Makhoul, F. Kubala, R. Leek, D. Lui, L. Nguqen, R. Schwartz and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE* vol. 88. no. 8, August 2000.

[16] C. Ng and J. Zobel, "Speech Retrieval using phonemes with error correction," *SIGIR* 1998.

[17] K. Ng and V.W. Zue, "Subword unit representations for spoken document retrieval," *Eurospeech* 1997.

[18] K. Ng and V.W. Zue. "Phonetic recognition for spoken document," *ICASSP* 1998.

[19] K. Ng. "Towards an integrated approach for spoken document retrieval," *ICSLP* 2000.

[20] T. Pfau, M. Beham, G. Ruske, "Creating large subword units for speech recognition," *Eurospeech* 1997.

[21] M. Siegler and M. Witbrock, "Improving the suitability of imperfect transcriptions for information retrieval from spoken documents, " *ICASSP* 1999.

[22] A. Singhal and F. Pereira, "Document expansion for speech retrieval," *SIGIR* 1999.

[23] K. Sparck Jones, G.J.F. Jones, J.T. Foote and S.J. Young, "Experiments in Spoken Document Retrieval," *Information Processing and Management* vol. 32 no. 4, 1996.

[24] H. Wang, H. Meng, P. Schone, B. Chen and W. Lo, "Multi-scale audio indexing for translingual spoken document retrieval," *ICASSP* 2001.

[25] V. Warnke, S. Harbeck, E. Nöth, H. Niemann, "Topic spotting using subword units," *9. Aachener Kolloquiem Signaltheorie, Bild- und Sprachsignale*, 1997.

[26] M. Wechsler, E. Munteanu, P. Schäuble. "New Techniques for Open-Vocabulary Spoken Document Retrieval". *SIGIR* 98.

[27] S.J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.-L. Gauvin, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual large vocabulary speech recognition the European SQUALE project," *Computer Speech and Language* 11, 1997.

[28] S.J. Young, et. al., "Acoustic indexing for multimedia retrieval and browsing," ICASSP 97.