Syllable-based Language Models in Speech Recognition for English Spoken Document Retrieval

Christian Schrumpf, Martha Larson, and Stefan Eickeler

Fraunhofer Institute for Media Communication, Sankt Augustin, Germany
martha.larson@imk.fraunhofer.de,
WWW home page: http://www.imk.fraunhofer.de

Abstract. The spoken content of audio/visual collections such as TV or radio archives is an information resource of enormous potential. The challenge is to develop methods that will make it possible to browse or search these collections. The experimental results presented in this paper demonstrate that syllable-level transcripts provide an important supplement to conventional word-level transcripts for the task of unlimited vocabulary American English spoken document retrieval. Recognition is performed with syllable language models with vocabulary sizes 20k, 10k, 5k, 1k, and 500. The syllable recognition rates of the 10k and 5k models are comparable to that achieved by a baseline 100k word-based language model. A simple retrieval experiment involving a fuzzy full text search supplies proof-of-concept that syllable-based transcripts make it possible to retrieve spoken documents that contain query words not included in the 100k vocabulary of the word-based language model.

1 Introduction

At present, the amount of multimedia data stored in digital collections, such as TV or radio archives, is already staggering. Without doubt, the trend of rapid accumulation of sparsely annotated multimedia data will continue into the future. The usefulness of a multimedia collection is a function of the ease with which it can be searched or browsed. A digital archive from which it is not possible to efficiently retrieve specific content is a stockpile of unexploited potential. The sheer quantities of archived multimedia material in existence generate pressing demand for methods that make multimedia data as easily accessible and searchable as text. The work reported on in this paper focuses on speech recognition methods for audio and video documents containing American English spoken content. In particular we investigate radio news. Speech recognition technology makes it possible to automatically produce transcripts from which indexing features can be extracted for spoken content. In order for spoken document retrieval to be maximally effective, the indexing features generated from spoken audio must not

only be reliable and discriminating, they must also comprehensively represent the spoken content of the audio/video document.

A persistent challenge for spoken document retrieval is the problem of Out Of Vocabulary (OOV) words, namely, words that do not occur in the vocabulary of the speech recognizer. The OOV problem arises from the fact that speech recognizer vocabularies, although large, are finite. Even though it is possible to automatically update vocabularies of speech recognizers as new or previously infrequent word forms come into regular use, such updates affect only new material entering the archive. Material already in the archive must be re-processed by the speech recognizer if its indexing is to reflect an update in the speech recognizer vocabulary. Although it is certain that size of the vocabulary which it is possible to use for speech recognition will increase, such increases will not serve to completely eliminate this problem. It is not possible to define a finite set of proper names and guarantee that only these names will ever gain prominence in the news. Instead, the features used to index spoken documents must be modified to encode proper names that are yet unknown at the time that audio documents are recognized and entered into the archive. In this paper we investigate such a modification.

The modification we chose to investigate is using syllables as indexing features for spoken audio documents as an alternative to orthographic words. Syllables are smaller than words and a relatively modest inventory of syllables provides building blocks that can be strung together to generate a nearly unlimited number of words. The potential of syllable-based spoken document retrieval is intuitively plausible for syllable-based languages like Chinese. The discriminative ability of syllables has been shown for Cantonese [1] and Mandarin [2]. Although German is not a syllable-based language, it is also an obvious candidate for the application of syllable methods. German includes full inflectional paradigms and makes liberal use of compounding to express concepts for which a language such as English would prefer to use a phrase. Because of these characteristics, even speech recognizers with very large vocabularies exclude numerous potential word forms. Due to inflection and compounding, many distinct word forms in German share the same syllables. For this reason, syllable-based transcripts suggest themselves as more compact representations of German spoken language content. In previous work we have demonstrated that syllables in German are useful indexing features for spoken document classification [3] and retrieval [4]. The goal of the work reported on here is to test the potential of syllable-based language models for speech recognition. We were encouraged by recognition results of a syllable-based English-language system reported in [5], which incorporated syllable acoustic models. Our aim is to achieve the benefits of syllable-based decoding without the effort of training syllable acoustic models.

In order to determine if the benefit of syllable-level features for spoken document indexing extends to English-language spoken document retrieval we implement and test a syllable based speech recognition system and perform exploratory spoken document retrieval experiments. Upon first consideration, English seems to be unsuitable for syllable-based recognition because it lacks full

inflectional paradigms and does not make extensive use of word compounding. Further consideration turns up additional factors, however, that suggested to us that it would be worth while to determine if syllable-based approaches are useful for English-language spoken document retrieval. First, syllable-based approaches promise to alleviate the OOV-problem. A syllable-based speech recognizer recognizes a word as a string of common syllables, whereas a word-based speech recognizer must have the full word form in its vocabulary, in order to be able to recognize it. Second, syllable-based approaches promise to help compensate recognizer error. A brief example serves to illustrate this point. In a case in which *Beslan* is included in the recognizer vocabulary, it still must be correctly recognized if it is to be used as a feature to index an audio document. Mis-recognitions of *Beslan* might include *best LAN* or *Bess line*. A standard search would not return a document indexed with either of these recognition mistakes to a user who is searching the archive with the query word *Beslan*. It is evident, however, that there is a similarity between the word spoken and the mis-recognized word substituted in the speech recognizer transcript. Representing words at the syllable level makes it possible to capture this similarity. The syllable representations of the mis-recognitions *best LAN* or *Bess line* are *best-lan* and *bes-layn*. Each of these representations has a syllable in common with the query word syllable string *bes-lan*. Moreover, *best* deviates in two phonemes from *bas* and *layn* deviates in one phoneme only from *layn*. By choosing an appropriate metric, we can represent the degree of similarity in such near matches. This example is idealized, but illustrates the point that syllable representations have the potential of error compensation for English language spoken document retrieval.

Section 2 discusses syllable-based speech recognition. The data corpus and the speech recognizer used for the experiments are introduced. We describe the training of the syllable-based language models and the word-based model used as a baseline. Then, we present the results of the speech recognition experiments. Section 3 discusses exploratory spoken document retrieval experiments performed to assess the potential of syllable-based indexing features. We introduce the fuzzy match method that we use to find near matches between query terms and syllable indexing features that signal a probable hit in the spoken audio. Then, we present the results of the spoken document retrieval experiments. Section 4 provides conclusions and sketches the direction that future work on syllable-based indexing features must take.

2 Syllable-based speech recognition

The speech recognition system used for the experiments was trained and tested on material from the Boston University Radio Speech Corpus (BURSC) published by the Linguistic Data Consortium in 1996. This corpus consists of 1666 short segments of radio news read by professional speakers and recorded at WBUR radio station. The speakers, three females and four males, are native speakers of American English. The recordings are of studio quality and con-

tain no background noise or sound effects. For this reason, they offer optimal conditions for speech recognition. The audio files are in mono NIST SPHERE format sampled with 16 kHz and were converted into WAVE format without any change in quality. For the experiments performed here, the corpus was divided into three parts. We designated 80% as the training set, 10% as the development set and 10% of the corpus as the test set. Section 2.1 describes the speech recognition system used for the experiments and the training of the acoustic models. Section 2.2 describes the training of the syllable-based language model, the innovative component of the speech recognition system. The training of the language model necessitates a large corpus of text and is independent of the training of the acoustic models. Section 2.3 discusses the speech recognition experiments and their results. Further details on the syllable-based recognition system and on experiments performed on the BURSC corpus can be found in [6].

2.1 Speech recognition system

For the experiments a classic Hidden Markov Model (HMM)-based speech recognition system is used which is implemented with the ISIP public domain ASR Prototype System 5.14 [7]. We use an inventory of 39 phonemes to represent the basic sounds of American English. For each phoneme we train a five-state HMM, iterating the training process until each density had 16 Gaussian mixtures. We train cross-word-triphone acoustic models using a nine-hour subset of the training set we had defined on the BURSC corpus. The nine hours of audio comprise 1325 files containing 5178 sentences, are used for the training of the acoustic models. Optimal parameter settings (language model weight and word start-up penalty) are empirically determined on the development set and then applied in decoding the test set.

2.2 Training the language models for the speech recognition system

To train the language models we combine a set of text corpora consisting in total of about 330 million words. The deployed corpora are the Reuters-21578 corpus (2.7 million words) and the TIPSTER Complete, Text Research Collection published by the Linguistic Data Consortium (LDC). TIPSTER Complete consists of the news wire of the Associated Press from 1988-1990 (107 million words), the news wire of the Wall Street Journal from 1987-1992, (69 million words), scientific texts of the Department of Energy containing (28 million words), news wire of San Jose Mercury News (30 million words) and *Computer select disks* of Ziff-Davis (92 million words). With this big text corpus we hope to get a distribution of the word types typical for news media. To use as a baseline, we train two word-based language models on this text. We chose 100k as the vocabulary size, since this is a size representative of large vocabulary word-based recognizers in common use. These 100k word forms include all word forms that occurred 28 or more times in the training text (see Table 1). We also trained a 10k word-based language model in order to provide a same-size performance comparison with a 10k syllable model. The 10k word model includes all forms

Table 1. Frequency with which word forms included in vocabulary of word-based language models occurred in the training text

Language model	Frequency cutoff for vocabulary
100k words	$\geq 28\times$
10k words	$\geq 2242\times$

occurring 2242 times or more in the training text. To train the language models we use the SRI language modeling toolkit [8]. We train two versions of each language model, a bigram version and a trigram version. Our experience is that Good-Turing smoothing, a common choice, is appropriate for both word and syllable language models. In order to be able to interface with the acoustic models representing the individual phonemes, a language model requires a pronunciation dictionary that specifies a string of phonemes for each word. We create the phonemizations necessary for the pronunciation dictionary using the text-to-phoneme software addttp4-1.1 [9] of the National Institute of Standards and Technology (NIST). We train five experimental syllable-based models with vocabulary sizes 20k, 10k, 5k, 1k and 500. The syllable models are trained on the same training text as the word models. In order to train a syllable model, it is necessary to first decompose each word of the text into its component syllables. We decompose the word-based text into the syllable-based text using the NIST syllabification software tsylb2 [10]. This decomposition process yields a version of the text consisting of one long string of syllables each separated with white space. The syllable version of the training text contained about 570 million running syllable forms. Table 2 summarizes how frequently syllables had to appear

Table 2. Frequency with which syllable forms included in vocabulary occurred in the training text

Language model	Frequency cutoff for vocabulary
20k syllables	$\geq 5\times$
10k syllables	$\geq 138\times$
5k syllables	$\geq 1741\times$
1k syllables	$\geq 70758\times$
500 syllables	$\geq 173881\times$

in the syllable version of the text in order to be included in the different syllable vocabularies. We test a syllable model with a 5k vocabulary since previous work at IMK has shown that for German syllable-based language models, recognition

rate improvement levels off at a syllable vocabulary size of 5k [11]. We chose a range of larger and smaller models in order to determine the effects of vocabulary size on recognition performance.

2.3 Results of speech recognition experiments

Speech recognition tests were performed on the test set consisting of 145 broadcast news segments (10% of the BURSC corpus) containing a total of 603 sentences and amounting to 68 minutes. Table 3 summarizes the recognition performance for the word-based language models. The recognition rates are reported

Table 3. Recogintion performance of word-based language models (baseline models)

Language model		Syllable accuracy	Decoding time
100k word	bigram	87.0 %	20.56 ×RT
	trigram	87.7 %	30.07 ×RT
10k word	bigram	81.4 %	10.49 ×RT
	trigram	81.7 %	11.59 ×RT

in terms of syllables in order to be comparable to the recognition rates of the syllable-based language model. If the recognition rates were reported in terms of words, they would be lower. For example, the best performance, which was achieved by the 100k vocabulary trigram language model, is a syllable recognition rate of 87.7%. This syllable rate corresponds to a word recognition rate of 80.5%. It can be seen that the time needed for recognition is significantly longer for the language models with the larger vocabulary. Of course, it is possible to reduce recognition times by decoding simultaneously on several processors. In some cases, for examples for mobile applications, the time needed to decode spoken audio using a single processor is relevant to the choice of the appropriate language model. Notice that the trigram models outperform the bigram models, but by a relatively slender margin. We believe that this fact is due to a mismatch between the training data used for the language models, and the test data. The training data simply did not contain many trigrams that matched trigrams occurring in the test data, and so moving from the bigram to trigram language model did not lead to a significant improvement.

The recognition performance for the syllable-based language model is summarized in Table 4. These experiments show that syllable-based language models with very small inventories (1k and 500 syllables) do not yield satisfactory syllable recognition rates. The 5k, 10k and 20k trigram syllable-based models perform well, and even exceed the syllable recognition rates achieved by the 100k word-based language model. These results support the conclusion that if both recognition performance and decoding time are to be taken into consideration,

Table 4. Recognition performance of syllable-based language models (experimental models)

Language model		Syllable accuracy	Decoding time
20k syllable	bigram	81.9 %	14.84 ×RT
	trigram	88.8 %	17.91 ×RT
10k syllable	bigram	81.8 %	13.51 ×RT
	trigram	88.8 %	15.54 ×RT
5k syllable	bigram	81.5 %	12.26 ×RT
	trigram	88.4 %	13.77 ×RT
1k syllable	bigram	67.4 %	9.47 ×RT
	trigram	71.3 %	10.84 ×RT
500 syllable	bigram	51.4 %	8.80 ×RT
	trigram	53.1 %	10.69 ×RT

the 5k trigram syllable-based model is the best choice and can be used instead of a 100k word-based model with a drastic enhancement in decoding performance.

3 Exploring syllable indexing features for spoken document retrieval

The performance of the syllable-based recognizer proves to be highly satisfactory. Syllable-based transcripts, however, cannot generally be read by humans and for this reason are not useful, unless they serve some specific purpose. In our case, we would like to use syllable-based transcripts automatically generated by the recognizer to extract syllable level indexing features that will be used for spoken document retrieval. In this section, we report the results of preliminary investigations performed in order to decide whether syllable indexing features merit pursuit for English-language spoken audio documents. Section 3.1 describes the fuzzy matching procedure that we use to find matches between user queries and strings of syllables occurring in the syllable transcripts. Section 3.2 discusses the experiments performed and their results.

3.1 Fuzzy match using syllable-level indexing features

The fuzzy match procedure identifies points in the syllable transcripts in which the syllable sequence generated by the speech recognition provides a close match to a query word submitted by a user. Our fuzzy match procedure determines a word-match score between the query word and every syllable string (sequence) in the syllable transcript. The word-match score is the Levenshtein distance between the syllable string composing the query word and a syllable string in the syllable transcript. This distance between syllable strings is weighted with

a syllable-match score for each syllable pair matched. The syllable-match score is the Levenshtein distance between syllables. We encode knowledge concerning the types of mistakes that the speech recognizer generally makes by weighting this distance with information concerning common phoneme confusions. The phoneme confusions are calculated on the development set. The document containing a syllable string with the highest word-match score is returned as the top ranking document in the query answer list. In order to constrain the length of the answer list it is necessary to either set a predefined length for the answer list, or to set a word-match score threshold and return only documents that exceed this threshold.

3.2 Experiments and results for spoken document retrieval

The fuzzy word match we introduce here uses the syllable output of the recognizer and further aims to exploit prior knowledge of recognizer error to find the spoken document relevant to the query. The spoken document retrieval experiments performed on the output of the speech recognizer are exploratory in nature and were formulated in order to help us determine whether or not syllable-based features merit further pursuit for English-language spoken document recognition. We performed two experiments each involving a different set of queries (search words) on the same test set used for the speech recognition experiments (145 broadcast news segments from the BURSC corpus). This section reports the results of these experiments.

In the first experiment, we decided to use each word token in the reference transcript as a one-word query. We used a modified version of the stopword list built by Gerard Salton and Chris Buckley for the SMART information retrieval system [12] in order to eliminate stopwords from the list of tokens we used as queries. The modifications we made to the stopword list involved mainly retaining numbers, which are important for dates, and retaining letters of the alphabet, which are important for abbreviations. We arrived at a list of 7242 queries.

Table 5. Experiment 1: Queries consisting of words in reference transcripts (stopwords removed)

transcript	exact	fuzzy match (thresh.)		
	match	0.70	0.65	0.60
word transcripts	86 %	-	-	-
words split into syllables	-	51 %	61 %	80 %
syllable transcripts	-	50 %	61 %	79 %
average length of list	-	23	28	33

In one test we used syllable transcripts generated by decomposing the word output of the word speech recognizer into syllables and in a second test we used syllable transcripts generated by the syllable speech recognizer. The retrieval performance at three threshold levels is reported in Table 5. The fuzzy match threshold controls the length of the hitlist returned for a query. The average length of the list is given for each threshold level. At level 0.60 the histlist has an average length of 33. If a hitlist of length 33 is randomly chosen from our collection of 145 documents there is a 23% chance it will contain the target document. Threshold levels higher than 0.60 have even higher random chance performance and we do not consider them interesting.

Performance on output from the syllable-recognizer (syllable transcripts) is approximately equivalent to performance on output from the word-recognizer that has been decomposed into syllables. Exact match on the word transcripts outperforms the fuzzy syllable match. This fact suggests that the most effective use of syllable indexing features is as a supplement to word indexing features. We performed a further experiment to test the syllable transcripts in cases where the word transcripts failed to contain the target word.

Table 6 summarizes the result of the second experiment in which we used as queries 990 word tokens that were not present in the output of the word-based speech recognizer. These queries are exactly the queries for which syllable-based indexing features represent critical supplement to word-based features. The results show 56% of the query words not found in the word-based transcripts can be found in the syllable-based transcripts using the fuzzy method and assuming that the user has the patience to consider a list of the 22 top fuzzy matches. Output from the syllable recognizer yields better results than output from the word recognizer decomposed into syllables.

Table 6. Experiment 2: Queries consisting of words in reference transcripts (stopwords removed) but not in speech recognizer output

transcript	exact	fuzzy match (thresh.)		
	match	0.70	0.65	0.60
word transcripts	86 %	-	-	-
words split into syllables	-	26 %	38 %	49 %
syllable transcripts	-	32 %	43 %	56 %
average length of list	-	12	17	22

4 Conclusions and outlook

The experiments reported on here have shown that a speech recognizer using a syllable-based language model of only a fraction of the size of a large word-

based language model achieves an equal or better syllable recognition rate. A 5k trigram syllable model yields a syllable accuracy approximately equivalent to that yielded by a 100k trigram word model, but the recognizer processes speech input more than twice as fast when using the syllable model than when using the word model. Further, exploratory spoken document retrieval experiments have led to the conclusion that syllable indexing features for spoken document retrieval are important as a supplement to word features. Syllable features can be created by decomposing word transcripts, so it is not strictly speaking necessary to decode once using a word model and once a syllable model. However, if only the query words not found in the word-recognizer output are considered, retrieval performance is better on syllable recognizer output than on decomposed word recognizer output. In future work, we would like to explore syllable features for English-language retrieval using a larger corpus. Additionally, it would be interesting to investigate whether performance can be enhanced by adapting the word-based language model for use in combination with syllable-based model for generation of features optimal for fuzzy word matching.

References

1. Meng, H. M., Lo, W. K., Li, Y. C., Ching, P. C.: Multiscale Audio Indexing for Chinese Spoken Document Retrieval. *Proceedings of ICSLP* (2000).
2. Chen, B., Wang, H., Lee, L.: Discriminating Capabilities of Syllable-based Features and Approaches of utilizing them for Voice Retrieval of Speech Information in Mandarin Chinese. *IEEE Transactions on Speech and Audio Processing* (2002) Vol. 10, Nr. 5.0.
3. Larson, M., Eickeler, S., Paass, G., Leopold, E., Kindermann, J.: Exploring subword Features and Linear Support Vector Machines for German Spoken Document Classification. *Proceedings of ICSLP* (2002).
4. Larson, M., Eickeler, S.: Using Syllable-based Indexing Features and Language Models to improve German Spoken Document Retrieval. *Proceedings of Eurospeech. 8th European Conference on Speech Communication and Technology* (2003).
5. Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G., Picone, J.: Syllable-Based Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing.* vol. 9, no. 4. (2001).
6. Schrumpf, C.: *Entwicklung und Evaluation eines silbenbasierten Spracherkenners fuer die englische Sprache.* Diplomarbeit, Fachhochschule Kaiserslautern/Fraunhofer IMK (2004).
7. Ganapathiraju, A., Deshmukh, N., Zhao, J., Zhang, X., Wu, Y., Hamaker, J., Picone, J.: *The ISIP Public Domain Decoder for Large Vocabulary Conversational Speech Recognition.* http://www.isip.msstate.edu (1999).
8. Stolke, A.: SRILM – An extensible language modeling toolkit. *Proceedings of the International Conference on Spoken Language Processing* (2002).
9. Fisher, B.: addttp4-1.1, http://www.nist.gov/speech/tools/index.htm NIST (2000).
10. Fisher, B.: tsylb2, http://www.nist.gov/speech/tools/index.htm NIST (1996).
11. Larson, M., Eickeler, S., Biatov, K., Koehler, J.: Mixed-unit language models for German language automatic speech recognition. *Proceedings of 13. Konferenz Elektronische Sprachsignalverarbeitung* (2002).
12. Salton, G.: *The SMART retrieval system: Experiments in automated document processing.* New Jersey: Prentice Hall (1971).