

Application of phonological knowledge in audio word spotting

Wai Yat Wong, John Robertson
CSIRO Mathematical and Information Science
Locked Bag 17, N Ryde, NSW 2113
Australia

wai.wong@cmis.csiro.au, john.robertson@cmis.csiro.au

Abstract

This paper describes the use of phonological knowledge to enhance current audio word spotting technology based on Hidden Markov Model. It reports on the result of an word-spotting experiment done on an audio data recorded from a live conference. The result of the experiment demonstrates the positive contribution of phonological knowledge towards audio word-spotting.

Introduction

This work is part of the Australian Vice Chancellor Committee (AVCC) Electronic Proceedings Project. The AVCC project is a hypermedia authoring research and development project with an objective to evaluate new models and techniques for authoring hypermedia titles. This hypermedia indexing application domain gives us the motivational backdrop to study how effective current automatic speech recognition technologies are for searching keywords within an audio stream.

Audio word spotting is the art of locating a query keyword directly within a speech signal stream without prior manual transcription. Hidden Markov Model (HMM) has been used extensively and successfully for word spotting (Rose and Paul 1990). However, existing techniques using HMM's have proven to be inadequate to manage the recognition tasks for speech data produced in real-life noisy environment where speaker independence and large open vocabulary are essential criteria (Woodland et al. 1997, Young et al. 1997a). This paper explores the application of high-level phonological knowledge on the word spotting problem to enhance the audio search result. It will begin by describing a brief overview of automatic speech recognition (ASR) technologies, a description of the audio word spotting problem domain, our approaches to deal with the word spotting problem, some experimental results, further works, and the conclusion.

History and relationship of AI and Automatic Speech Recognition (ASR)

There has been a lot of research and development of

automatic speech recognition by machine for half a century. Ainsworth in his book, "speech recognition by machine" (Ainsworth 1988) gives a very good overview of the various approaches researchers had tried through the years. By the late eighties and early nineties, speech research activities had concentrated on sophisticated mathematical and computing techniques like dynamic programming, stochastic models, artificial neural net, Hidden Markov Model (HMM), and fuzzy-set (Ainsworth 1988, Rabiner, Juang, and Lee 1996). These mathematical and computing techniques have contributed significantly to the ASR but recent papers seemed to indicate a performance plateau (Woodland et al. 1997, Young et al. 1997a, Gibbs 1997). Interest has re-emerged for the integration of domain specific and ad hoc high level domain knowledge, linguistic knowledge, and perhaps other new AI techniques back to the ASR domain. It is within this framework that we are exploring the use of general phonological knowledge on top of the current technological dominant HMM model to improve the performance on audio word spotting.

Problem domain – audio word spotting

The objective of the audio word spotting task is to identify the boundaries of a search term within a digitised continuous speech stream. Because our objective is to search and index real-world commercial multimedia archives, the audio data to be managed is particularly problematic for the current capabilities of automatic speech recognition technology. The automatic speech recognition system must be designed to manage the following data characteristics:

Speaker independence: Media archives typically contain data from more than one speaker. Because the speech recognition systems is trained from a specific group of representative speaker, the set of new speakers that can be included in the search archive can possess different attributes, e.g. gender, age, language and dialect variances, which will impact on the performance of the automatic speech recognition system.

High noise level: Because there is no guarantee that the data in the archive has been captured in an environment that eliminates noise, the system must be able to manage background noise. The performance levels of most current speech recognition methodologies degrade

significantly when there is environmental noise. Our experimental data was recorded in an auditorium with background audience noise and additional noise from the recording equipment.

Open vocabulary: We do not make any assumption on the vocabulary of the query keyword term and the target audio archive data for searching. This is contrasted with some speech recognition systems which requires that the query keyword and the target audio archive data for searching must be bounded by a controlled set of vocabulary. Furthermore, we should be able to cater for any query keyword that can be sufficiently represented by an arbitrary combination of standard linguistic phonetic set (e.g. International Phonetic Alphabet (IPA) (Akmajian, Demmers, and Harnish 1987, Macquarie 1990)).

Continuous speech: The audio data will contain data spoken in a normal conversational manner. This is contrasted to a lot of studies where the speech recognition is done on isolated word or finite set of isolated phrases predefined and constrained by a grammar.

Experimental approach

The underlying speech processing engine for our audio word spotting exercise is based on the HTK toolkit (Young et al. 1997b) from the Entropic Research Laboratory for building hidden Markov Model based speech recognisers.

The process consists of three sequential phases: the HMM training phase, the grammar specification phase, and the recognition/word spotting stage. Figure 1 below depicts the process.

HMM training phase: building required HMM via training

For the training phase, we use an ANDOSL speech corpus (Miller et al. 1994). This database contains a collection of sentences spoken by Australians and has been previously acoustic-phonetic-labelled. Acoustic-phonetic labelling involves segmenting the speech signal in terms of phonetic characteristics. In principle, segments which have the same acoustic-phonetic characteristics are given the same label, and each assigned label links the acoustic-phonetic segment to the phonetic entity it represents. The set of labels associated with a speech signal file constitutes a transcription and each transcription is stored in a separate label file. In our experiments, continuous speech data of 200 sentences from 4 different speakers from the ANDOSL corpus were used as a training dataset.

The raw speech waveform for training is parameterized into speech vectors. Based on the standard DARPA TIMIT Acoustic-Phonetic continuous Speech Corpus phone set (Lamel, Kassel, and Seneff 1996) and the International Phonetic Alphabet (Macquarie 1990), we derive a reduced set of 45 phones (figure 2) and build a HMM for each phone; the 45 HMM's include a silence and pause model. Together with a prototype HMM definition, the speech vectors and labels are fed into a training phase to estimate the parameters for the HMM's. The algorithms used in the training phase are the Baum-Welch re-estimation and the forward-backward algorithm (Young et al. 1997b). The output of this training phase is the required set of HMM phone models used for the word spotting exercise.

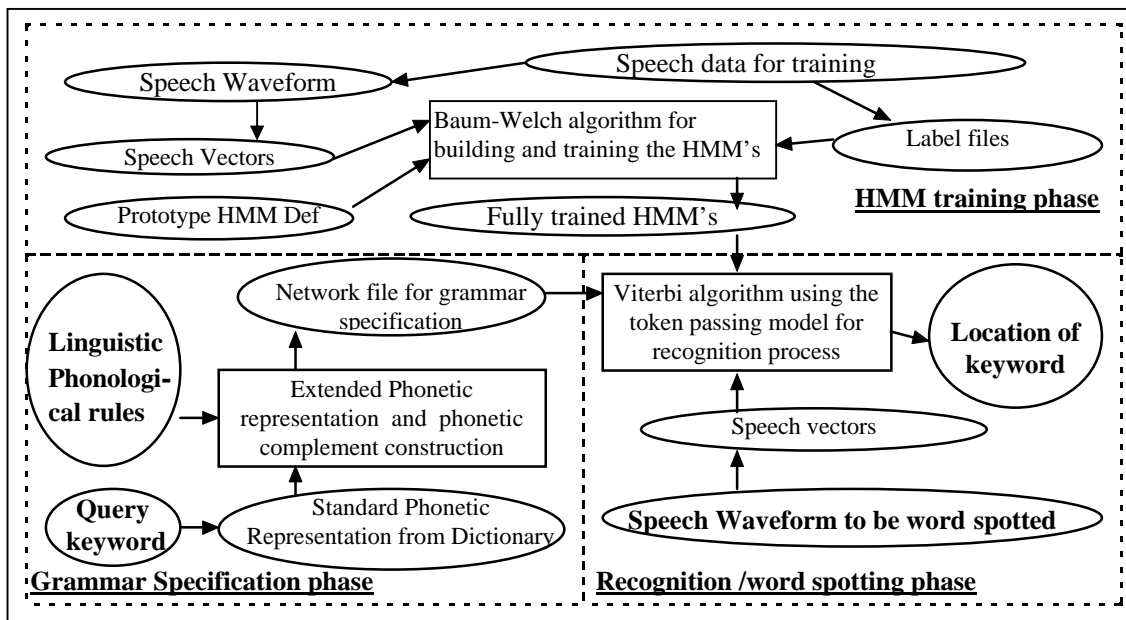


Figure 1: The Experimental Process

Phone groups	Phones
Stops(Plosives)	p, t, k, b, d, g
A/Fricatives	T, D, f, h, v, s, z, S, Z
Nasals	m, n, N
Approximants (Semivowels)	w, l, r, j
Short vowels	V, E, A, I, O, U, @
Long vowels	a: , e: , i: , o: , u: , @:
Diphthongs	ei, @u, oi, ai, au, I@, u@
Other	H#, #, J

figure 2: phone set

Grammar Specification Phase : building the grammar necessary for word spotting

The query keyword is mapped to a standard phonetic grammar representation string using a standard phonetic dictionary (Macquarie 1990). The resultant phonetic representation is extended using phonological knowledge from the linguistic research to allow for other possible variant pronunciation under general context. The grammar for the anti-phonetic pattern which forms the complement of the extended phonetic representation string for the query keyword is also constructed. Both the grammars for the phonetic string of the query keyword and its complementary anti-phonetic pattern, are fed into the grammar specification network file.

This section illustrates an example with the query keyword “library”. Using standard phonetic dictionary, a query keyword like “library” would be converted to a phonetic stream like /l ai b r @ r i/ using a standard Australian phonetic dictionary. Applying the phonological rules, the phonetic stream is then extended to be / (r)l (@|ai) (v)b [r] @ r (i | I)/. The grammar for the phonetic complements of the keyword is also constructed by the anti-word grammar generation module.

The combination of the valid phonetic grammar representations of the keyword and the corresponding grammar for the phonetic complements will produce grammar specification network as shown below:

/* | = or, [] = optional, concatenation = and,
compl. = complement */

```
$LIBRARY = WD_BEGIN% LIBRARY (r)l (@|ai) (v)b
[r] [ @ ] (r)l (i:I) WD_END% LIBRARY;
$ff1 = f | h | v | T | D | s | S | z | Z;
$ff1v = f | h | T | D | s | S | z | Z; /* Fricatives & compl. */
$pp1 = p | t | k | b | d | g;
$pp1v = p | t | k | d | g; /*plosive & compl.*/
$nn1 = m | n | N; /* nasal & compl.*/
$vv1 = V | E | I | O | U | @ | A | a: | e: | i: | o: | u: | @: | ei
| @u | oi | ai | au | i@ | u@;
$vv2 = V | E | I | O | U | A | a: | e: | i: | o: | u: | @: | ei |
@u | oi | au | i@ | u@; /* vowel & diphthongs & compl */
$vv3 = V | E | I | O | U | A | a: | e: | i: | o: | u: | @: | ei |
@u | oi | ai | au | i@ | u@;
$vv4 = V | E | O | U | @ | A | a: | e: | o: | u: | @: | ei | @u
```

```
| oi | ai | au | i@ | u@;
$aa1 = w | l | r | j; $aa2 = w | j; /* semivowel & compl. */
$oo1 = H# | # | J;
$g0 = $aa2 | $pp1 | $nn1 | $oo1 | $ff1;
$g1 = $ff1v | $pp1v | $nn1 | $oo1 | $aa1;
$Garbage = WD_BEGIN% Garbage $g0 $vv2 $g1 $vv3
$g0 $vv4 WD_END% Garbage;
( <$LIBRARY |$Garbage> )
```

Here is a brief explanation of how phonological knowledge can be used during the grammar network specification stage. Phonological rules express the systematic ways in which the pronunciation of words or phones may change with their environment. For example, vowel reduction is something common in casual spoken English. The word “potato”, if pronounced slowly, will have this phonetic representation /p @ t ei t @u/. But in continuous casual speech, the first vowel, @, can sometimes be elided. Thus, the phonetic representation of “potato” could become /p t ei t @u/(Hawkins 1984). So the phonological rule can produce an extended phonetic representation for “potato” to be /p [@] t ei t @u/ (where [] means ‘optional’). Similarly, in continuous speech, there is the phenomenon of consonant deletion, e.g., the sequence “six seven” is normally pronounced as /s I k s E v n/ or /s I s E v n/ instead of /s I k s s E v n/ (Ainsworth 1988); or, the dropping of consonant h in the sequence “twist his arm” to produce /t w I s t I z a: m/ instead of /t w I s t h I z a: m/(Hawkins 1984). Their extended phonetic representation will respectively be /s I [k] s [s] E v n/ and /t w I s t [h] I z a: m/. There is also Sandhi consonant insertion(Ainsworth 1988); the word “how” is pronounced as /h au/ and “are” is pronounced /a:/, but “how are you?” is pronounced as /h au w a: j u:/ with a phone /w/ is inserted in between. The extended phonetic representation for the query keyword “how are you” can thus become / h au [w] a: j u:/. Other rules include the possible substitution of accented vowel with reduced vowel /@/ (schwa) (Akmajian, Demmers, and Harnish 1987). For example, /E/ in “democrat” /d E m O k r A t I v/ is sometimes substituted with a reduced vowel /@/in “democracy” /d @ m O k r A t I v/. Thus to be inclusive, the extended phonetic representation can be represented as /d (E/@) m O k r A t I v/ for the query keyword “democrat”(where | means ‘either or’). Phonological rules may also be expanded to capture different ethnic background; for example, the word “three” with the phonetic representation of /T r i / can be often mispronounced as /t r i/ by a certain group of people. Again, in this case, the phonological module will produce an extended phonetic representation of /(T)t r i/.

Recognition /word spotting phase: labelling test speech data as word and anti word pattern

For the recognition/word spotting phase, the speech archive which forms the target space for searching is digitised and parameterised to speech vectors. The speech vectors will be input to a Viterbi module toolkit(Young et

al. 1997b) which uses the previously trained HMM's and the grammar specification network file to produce a resultant transcribed output. The Viterbi module makes use of the Viterbi algorithm with token passing. The resultant transcribed output will consist of locations of the query keyword and locations of the anti-keyword.

Below is a sample of the transcribed output from word spotting. The query keyword is "library". The number to the left is the end-time in second when the word label on the right has been spoken. The starting time of the label is taken to be the end-time of the previous label and 0 initially. Note that the output will contain either the keyword or the "anti-keyword" – Garbage word.

```
#
0.130000 121 LIBRARY
0.530000 121 Garbage
0.920000 121 LIBRARY
2.760000 121 Garbage
3.460000 121 Garbage
4.500000 121 Garbage
4.970000 121 Garbage
5.520000 121 Garbage
5.920000 121 Garbage
6.660000 121 Garbage
6.920000 121 Garbage
8.120000 121 Garbage
8.510000 121 Garbage
8.680000 121 LIBRARY
9.180000 121 Garbage
.....
```

Result and analysis

One of the popular performance measurement in speech community has been Figure of Merit(FOM) and the Receiver Operation Curve(ROC) (Young et al. 1997b). Since we believe that word spotting is more akin to information retrieval and the metrics used in information retrieval is intuitive and easily understood, we have dropped the complex FOM and ROC metrics and have opted for the metrics used in information retrieval community – recall and precision (Salton 1989).

recall = (number of retrieved relevant phrases via word spotting) / (total number of relevant phrases in the original audio stream)

precision = (number of retrieved relevant phrases via word spotting) / (total number of retrieved phrases via word spotting)

Regardless of which performance measurement used, one interesting behaviour is observed. In quite a few occasions the word spotting produces instances of retrieved phrase which appear wrong because the target query key-phrase does not show up in the same location on the search archive, but on closer examination, the retrieved phrases are in fact partial match – hence they are retrieved during the word spotting process. Examples : query string of "electronic publishing" versus the

retrieved phrase of "electronic communication". We are working on a satisfactory metric to accommodate partial match. Meanwhile, we do not count these partial matches as retrieved relevant phrases.

Below are two tables which record the recall and precision for some of the query phrases that we have used in the experiment. The result of the first table is produced without any use of phonological knowledge. The result of the second table is produced with extending the grammar using the phonological rules.

Without any use of phonological knowledge

Query keyword	Recall	Precision
"electronic publishing"	8/8	8/9
"Australia"	0/3	0/1
"library"	0/5	0/0
"today"	0/4	0/0

With the use of phonological knowledge

Query keyword	Recall	Precision
"electronic publishing"	8/8	8/10
"Australia"	3/3	3/50
"library"	3/5	3/15
"today"	3/4	3/21

The result seems to indicate that the expansion of the query keyword using phonological knowledge improves the performance of word spotting both in recall and precision. In view of the requirement that the environment could be noisy and the speech data can come from any speaker with an open vocabulary, the improvement is not surprising. The additional variant phonetic representation of the query keyword using the phonological rule will capture some of the possible ways the keyword is spoken by a general speaker and is perturbed by noise.

It is interesting to note that if the performance without phonological knowledge is reasonable as in the case for query keyword "electronic publishing", the use of phonological knowledge to extend the query keyword phonetic representation actually produce a worse precision value. Again, this is expected, the word spotter retrieves more than necessary.

Further work

Using more sophisticated HMM's

Our HMM's are trained using a small training dataset of eight hundred sentences from four individual speakers. Entropic Laboratory has some commercial pre-trained HMM's (Power et al. 1997) for speech recognition which use highly sophisticated HMM architectures & language models. These HMM's are also trained from very large dataset from large number of speakers and with very large vocabulary. We will conduct the same experiment using those HMM's. We expect the same positive contribution

from the application of phonological knowledge.

Better phonetic complement grammar model and approximate pattern matching approach

In the word & anti-word grammar model, the word spotting system has to construct all valid variant phonetic representations for the keyword (set P) and their phonetic complements (set P¹). The construction of the phonetic complements to the query keyword can be delicate. So far we have tried constructing the grammar for the keyword phonetic complements to be all possible phonetic combination minus the possible phonetic representation of the keyword.—the full complement method. We have also tried constructing the grammar for the keyword phonetic complements to be all possible combination of complement phones to the syllable level – the syllable complement method. The full complement has not produced a good result because of the noisy phonetic recognition. But the syllable complement has done better. The upshot of this is that we find this word and anti-word grammar model not as easy to use as we have hoped for.

We are also exploring the approximate pattern matching approach. In this approach, the audio stream is first recognised by our HMM's into a complete continuous phonetic stream. We assume this resultant phonetic stream to be error prone. We then use an approximate pattern-matching technique which is extended from "agrep"(Dept. of Computer Science, University of Arizona) to see if we can extract the phonetic representation of the query keyword from the phonetic stream. A confusion matrix produced from the HTK toolkit, which contains probabilities of one phone being confused with another, is being used together with phonological knowledge to reduce the search space.

Extend and complete the known linguistic phonological rules

We are collaborating with linguists to capture linguistic phonological rules into a complete phonological knowledge module and to develop a good algorithm for the rule application. Other useful linguistic knowledge which includes the study of stress patterns will also complement the phonological knowledge. Concepts from natural language processing can also be used to add new syntactic & lexical constraints and other semantic constraints.

Conclusion

The integration of AI based techniques with high-level linguistic knowledge on top of the current HMM based speech signal processing engine can bring about realistic ASR performance. The authors are exploring other alternative AI based techniques which can make effective use of the current pool of linguistic knowledge to achieve reasonable result for audio word spotting.

References

- (Ainsworth 1988)Ainsworth, W. A. 1988. *Speech Recognition by Machine*. Peter Peregrinus Ltd. London, United Kingdom.
- (Akmajian, Demmers, and Harnish 1987)Akmajian, A.; Demers, R. A.; and Harnish, R. M. 1987 *Linguistics: An Introduction to Language and Communication*. 2nd ed., MIT Press.
- (Gibbs 1997)Gibbs, W. W. 1997. Taking Computers to Task. *Scientific American*, July 1997, PP82-89.
- (Hawkins 1984)Hawkins P. 1984. *Introducing phonology*. Hutchinson, London.
- (Lamel, Kassel, and Seneff 1996)Lamel, L. F.; Kassel, R. H.; and Seneff, S. 1996. Speech Database Development: Design and Analysis of Acoustic Phonetic Corpus. In Proceeding of DARPA Universal Speech Recognition Workshop (editors Baumann, L.S.), PP100-109.
- (Macquarie 1990)Macquarie Dictionary Editorial Committee 1990. *The Macquarie Encyclopedic Dictionary*. The Macquarie Library Pty Ltd.
- (Miller et al. 1994)Miller, B.; Vonwiller, J.; Harrington, J.; and Dermody, P. 1994. The Australian National Database of Spoken Language. In ICASSP94.
- (Power et al. 1997)Power, K.; Matheson, C.; Ollason, D.; and Morton, R. 1997. *The graphHvite book*. Entropic Cambridge Research Laboratory Ltd, Cambridge University.
- (Rabiner, Juang, and Lee 1996)Rabiner, L.R; Juang, B. -H.; and Lee, C. -H. 1996. An Overview of Automatic speech Recognition. In *Automatic Speech and speaker Recognition: Advanced Topics* (editors Lee, C.-H., Soong, F.K., and Paliwal, K.K.) Kluwer Academic Publishers.
- (Rose and Paul 1990)Rose, R. C. and Paul, D.B., 1990. A hidden Markov model based keyword recognition systems. In Proc. Int. conf. on acoust., Speech and Sig. Processing, Apr 1990.
- (Salton 1989)Salton, G. 1989. *Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company.
- (Woodland et al. 1997)Woodland, P.; Gales, M.; Pye, D.; and Young S. 1997. Broadcast News Transcription using HTK. In ICASSP97, PP719-722
- (Young et al. 1997a)Young, S.; Brown, M.; Foote, J.; Jones, G.; and Jones, K. 1997. Acoustic Indexing for Multimedia Retrieval and Browsing, In ICASSP97, PP199-202.
- (Young et al. 1997b)Young, S.; Odell, J.; Ollason, D.; Valtchev, V.; and Woodland, P. 1997b. *The HTK Book*. Entropic Cambridge Research Laboratory Ltd, Cambridge University.