Phone recognition for Spoken Web Search

Etienne Barnard, Marelie Davel
Multilingual Speech Technologies, North-West University
Vanderbijlpark, South Africa
{etienne.barnard, marelie.davel}@gmail.com

Charl van Heerden, Neil Kleynhans
HLT Research Group,CSIR Meraka Institute
Pretoria, South Africa
{cvheerden, ntkleynhans}@gmail.com

Kalika Bali
Microsoft Research Lab India
Bangalore, India
kalikab@microsoft.com

ABSTRACT

Aiming at both speaker independence and robustness with respect to recognition errors in the spoken queries, we have implemented a two-pass system for spoken web search. In the first pass, unconstrained phone recognition of both the query terms and the content audio is employed to represent these recordings as phone strings. A dynamic-programming approach then finds regions in the content phone strings that correspond closely to one or more query strings. In the second pass, each of these regions is again processed with a phone recognizer, but now a lattice is extracted; this lattice is compared against similar lattices extracted for each of the queries. We find our approach to be somewhat successful in identifying the query terms in both the development and evaluation sets, but not to generalize well between these sets.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*

General Terms

Spoken term detection, under-resourced languages, confidence measures

1. INTRODUCTION

The 'spoken web search' task of MediaEval 2011 [4] involved searching for audio content in one of 4 under-resourced languages (Gujarati, Telugu, Hindi and Indian English), using only an audio version of the content query. All audio content [3] was collected over a mobile connection and acoustic quality varied. Our approach to this task was guided by three principles: (1) Since the search task requires speaker independence, we preferred to use standard speaker-independent ASR technology, rather than (say) template-based methods. (2) Any pronunciation model derived from a single spoken example of a search term is likely to be quite fragile; hence, specific care must be taken to model variability around such a model. (3) Only limited resources are available in the target languages / dialects; we therefore focused our efforts on approaches that did not rely on any textual data (or derived language models), and could produce results when closely matched ASR systems are not available.

2. APPROACH

As we did not have access to closely matched ASR systems for any of the target dialects, we focused our approach on obtaining data and building a set of acoustic models for at least one of the languages. These acoustic models were then adapted to the other languages and used during both spoken term detection and confidence scoring. Without time alignments for the development set, we generated our own using a grapheme-based system in order to evaluate our results during development.

2.1 Acoustic modeling

The main data set used for acoustic modelling was a Hindi corpus obtained from Microsoft. Several additional corpora were also considered, but were not as directly suited to this task.

2.1.1 Hindi Models

The Microsoft Hindi Corpus consists of 60 hours of spontaneous conversations in colloquial Hindi, recorded on Appen's telephony recording platform and sampled at 8kHz. There are 996 native Hindi speakers and all conversations range between 1 and 4 minutes in duration. All conversations are transcribed and time aligned on speaker turns, and a basic pronunciation dictionary is provided.

An initial acoustic model trained on all the audio was used to further process the corpus, using techniques described in [1]. The initial acoustic models trained were standard 3-state left to right tied triphone models, with 8 mixtures per state and semi tied transforms. A garbage model was then trained and combined with this initial model. The models were further refined by doing MAP adaptation using the target audio (in all 4 languages). Target audio transcriptions were generated using the initial acoustic models to decode the target audio, using a flat phone grammar.

The list of monophones was reduced rather aggressively from 62 to two smaller sets (of 43 and 21 monophones, respectively) in order to work with a small set of broad but reliable classes, appropriate to the later scoring tasks. The reduction included amongst others modeling aspiration separately, splitting diphthongs as well as affricates, combining some allophones and merging all the nasalized vowels with the corresponding non-nasal phonemes.

2.2 Spoken term detection

Spoken term detection was performed using a dynamic programming (DP) approach: audio data and query data are decoded by the ASR system using a flat phone-loop grammar, and the resulting phoneme strings matched against one another using a dynamic programming algorithm and a variable cost matrix. A phone set is used where transitional sounds (such as affricates or diphthongs) are split into their constituent parts. The phone string generated from the audio data is then segmented into detection candidates using a shifting window with a size matching the query phone string (plus or minus a leniency factor), and an alignment cost generated. The alignment cost, normalised by the phoneme length of the longer phone string is used directly as the DP score. This approach is influenced by both

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

the granularity of the phoneme set used and the scoring matrix. In this work a linguistically motivated matrix, a matrix derived from the posterior probabilities obtained from an ASR confusion matrix and a flat matrix were used.

Two additional approaches to detection were considered:

- *Grammar-based term detection*: A constrained decoding network grammar is constructured by placing repeatable phoneme fillers before and after the desired search term and allowing multiple search term detections within an utterance.
- *Lattice-based methods*: A phone lattice is constructed from the entire utterance within which search is to be performed, and the phoneme strings corresponding to each of the queries is matched against the lattice. However, the acoustic ambiguities of phone recognition in low-quality audio caused practical difficulties – both the computation required and the size of the resulting lattices were found to be unmanageable.

2.3 Confidence scoring

The *DP scores* generated during the detection phase can be used directly as confidence measures. In addition, two other confidence measures were calculated using the terms flagged during DP scoring: *lattice-to-lattice matching* and *dynamic time warping*. Since standard lattice-based confidence measures are difficult to utilise without reliable language models, a direct lattice-to-lattice matching measure was implemented. This is a direct extension of the DP-based string matching process, and can be implemented efficiently using an algorithm as described in [2], combined with similiar scoring matrices as for DP scoring. Posterior probabilities are obtained directly from the lattices, and a series of start and end points (relative to the initial detection) can be evaluated efficiently. Finally, Dynamic Time Warping (DTW) was used to match the query and detection on a frame-to-frame basis, and the corresponding DTW distance (normalized by number of frames) used as the confidence measure.

2.4 Evaluating results

To analyse our results, it was necessary to generate alignments for the development and evaluation data sets. Since the Indo-Aryan and Dravidian languages have a high letter-to-sound correlation it was decided to use a grapheme-based recognition system to generate alignments. The grapheme-based alignment system was represented by 8 mixture context-dependent tri-letter HMM acoustic models. A pronunciation dictionary was created by letter splitting the words, which resulted in 26 sub-word units (ASCII 'a' to 'z' in the English alphabet).

3. RESULTS

Our experimental results are based on detection of the development queries in the development data. Fig. 1 shows the standard DET curve obtained when using the confidence scores of the DP alignment to score hypotheses, as well as the DET curve when the lattice-to-lattice measure is employed. We see that the DP confidence scores yield better DET curves and OTWVs. A separate analysis showed that the lattice scores for valid matches are generally higher than those for false detections; we therefor searched for linear combinations between DP and lattice scores that would outperform either on its own. However, no consistent improvement was found. We therefore decided to use the DP scores in our submission. The ATWV scores obtained in the four task conditions are summarized below. These results confirm the fact that our system is somewhat successful in detecting the desired search terms; however, the negative ATWV scores across the dev/eval divide suggests that our system is quite sensitive to the differences between the two sets.

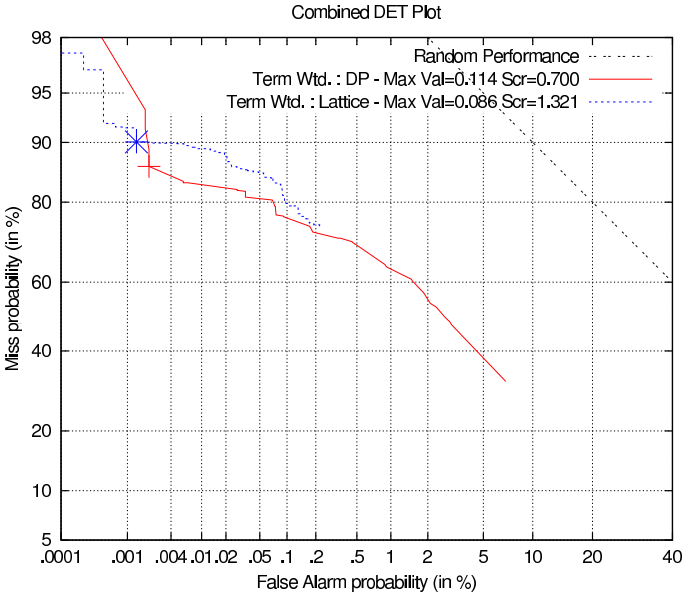

Figure 1: DET curves when using confidence scores based on DP alignment vs lattice-to-lattice posteriors.

Table 1: ATWV scores for four task conditions

Task	dev/dev	dev/eval	eval/dev	eval/eval
ATWV	0.102	-0.021	-0.13	0.114

4. CONCLUSION

We have presented a DP-based approach to spoken term detection, and a lattice-based scoring mechanism that was intended to refine the DP scores. The former approach was somewhat successful, but we have not been able to obtain a benefit from the latter; thus, our second assumption in Section 1 has not been confirmed. Given our time constraints, we have not been able to experiment systematically with numerous variables which are clearly important to the performance of both stages of the system (e.g. different phone sets, scoring matrices, lattice-extraction parameters, etc.) – it is likely that significant improvements within the current framework can be achieved by paying closer attention to each of these factors. Once that has been achieved, further score improvements by using additional acoustic front ends should be a straightforward (though computationally expensive) step.

5. REFERENCES

[1] M. H. Davel, C. van Heerden, N. Kleynhans, and E. Barnard. Efficient harvesting of Internet audio for resource-scarce ASR. In *Proc. Interspeech*, Florence, Italy, August 2011.

[2] Y. Kobayasji and Y. Niimi. Matching algorithms between a phonetic lattice and two types of templates – lattice and graph. In *Proc. ICASSP*, pages 1597 – 1600, Tampa, Florida, April 1985.

[3] A. Kumar, N. Rajput, D. Chakraborty, S. K. Agarwal, and A. A. Nanavati. WWTW: The World Wide Telecom Web. In *DSDR 2007 (SIGCOMM Workshop)*, Kyoto, Japan, August 2007.

[4] N. Rajput and F. Metze. Spoken Web Search. In *MediaEval 2011 Workshop*, Pisa, Italy, September 2011.