SWS task: Articulatory Phonetic Units and Sliding DTW

Gautam Varma Mantena
Speech and Vision Lab
International Institute of Information Technology, Hyderabad
Andhra Pradesh, India
gautam.mantena@research.iiit.ac.in

Bajibabu B
Speech and Vision Lab
International Institute of Information Technology, Hyderabad
Andhra Pradesh, India
bajibabu.b@research.iiit.ac.in

Kishore Prahallad
Speech and Vision Lab
International Institute of Information Technology, Hyderabad
Andhra Pradesh, India
kishore@iiit.ac.in

ABSTRACT

This paper describes the experiments conducted for spoken web search at MediaEval 2011 evaluations. The task consists of searching for audio segments within audio content using an audio query. The current approach uses a broad articulatory phonetic units for indexing the audio files and to obtain audio segments. Sliding DTW is applied on the audio segments to determine the time instants.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Spoken Term Detection, Articulatory Phonetic Units, Sliding Dynamic Time Warping

1. INTRODUCTION

The approach attempted aims at identifying audio segments within an audio content using an audio query. Language independency is one of the primary constraints for the spoken web search task [1]. We have implemented a two stage process in obtaining the most likely audio segments. Initially all the audio files are indexed based on their corresponding articulatory phonetic units. The input audio query is decoded into its corresponding articulatory phonetic units and necessary audio segments which have a similar sequence are selected. Sliding window type of approach using dynamic time warping (DTW) algorithm has been used in determining the time stamps within the audio segment. The following approach is provided in more detail in section 2.

2. TASK DESCRIPTION

A variant of DTW search is used to identify the audio segments within an audio content. DTW algorithm is time consuming and so it is necessary for the system to select the necessary segments beforehand. The system consists of an indexing step which would improve the retrieval time by selecting the required audio segment within the audio content. Following approach is a two level process where it prunes appropriate segments from each level. The two levels implemented are as follows:

1. First level: Indexing the audio data in terms of its articulatory phonetic units and using them to obtain the most likely segments for the input audio query.
2. Second level: Use of sliding window DTW search and k-means clustering to obtain the segments with the best scores.

The procedure for the audio search task is as described in sections 2.1, 2.2

2.1 Indexing using Articulatory Phonetic Units

The primary motivation for this approach is to have speech specific features rather than language specific features like phone models. Advantage is that the articulatory phonetic units (selected well) could represent a broader set of languages. This would enable us to build articulatory units from one language and use it for other languages.

Articulatory units selected are as shown in table 1. For example the articulatory unit *CON_VEL_UN* is a *consonant velar unvoiced* sound. A more detailed description of the tags is given in table 2.

Table 1: *Articulatory Phonetic Units derived from a Telugu database.*

Articulatory Unit	Phones
CON_VEL_UN	/k/, /kh/
CON_VEL_VO	/g/, /gh/
CON_PAL_UN	/ch/, /chh/
CON_PAL_VO	/j/, /jh/
CON_ALV_UN	/t:/, /t:h/
CON_ALV_VO	/d:/, /d:h/
CON_DEN_UN	/t/, /th/
CON_DEN_VO	/d/, /dh/
CON_BIL_UN	/p/, /ph/
CON_BIL_VO	/b/, /bh/
NASAL	/ng∼/, /nd∼/, /nj∼/, /n/, /m/
FRICATIVE	/f/, /h/, /h:/, /sh/, /shh/, /s/
VOW_LO_CEN	/a/, /aa/
VOW_HI_FRO	/i/, /ii/
VOW_HI_BAC	/u/, /uu/
VOW_MI_FRO	/e/, /ei/
VOW_MI_BAC	/o/, /oo/
/y/	/y/
/r/	/r/
/l/	/l/
/v/	/v/

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

Table 2: *Articulatory tags and their corresponding description.*

Articulatory Tag	Description
CON	Consonant
VEL	Velar
PAL	Palatal
ALV	Alveolar
DEN	Dental
BIL	Bilabial
VO	Voiced
UN	Unvoiced
VOW	Vowel
HI	High
MI	Mid
LO	Low
FRO	Front
CEN	Center
BAC	Back

Audio content is decoded into their corresponding articulatory units using HMM models with 64 Gaussian mixture models using the HTK Tool Kit [2]. The models were built using 15 hours of telephone Telugu data [3] consisting of 200 speakers. Using the decoded articulatory phonetic output, trigrams were used for indexing. The audio query was also decoded and the audio segments were selected, if there was a match in any of the trigrams. Let t_{start} and t_{end} are the start and the end time stamps for the trigram in the audio content which matches with one of the trigrams from the audio query. Then the likely segment from the audio content would be ($t_{start} - audio\ query\ length$) and ($t_{end} + audio\ query\ length$). This is would enable to capture speech segments with varying speaking rate.

These time stamps would provide the audio segments that are likely to contain the audio query. Sliding DTW search was applied on these audio segments to obtain the appropriate time stamps for the query which is explained in detail in section 2.2.

2.2 Sliding Window DTW Search

In a regular DTW algorithm the audio segments are assumed to have a timing difference and the algorithm helps in time normalization, i.e. it fixes the beginning and the end of the auido segments. In spoken term detection we also need to identify the right audio segment *within* an audio with appropriate time stamps.

We propose an approach where we consider an audio content segment of length twice the length of the audio query, and a DTW is performed. After a segment has been compared the window is moved by one feature shift and DTW search is computed again. MFCC features, with window length 20 msec and 10 msec window shift have been used to represent the speech signal. Consider an audio content segment S and an audio query Q. Construct a substitution matrix $\mathbf{M}$ of size q x s_q where q is the size of Q and $s_q = 2 * q$. We also define $\mathbf{M}$[i,j] as the node measuring the optimal alignment of the segments Q[1:i] and S[1:j]

During DTW search, at some instants $\mathbf{M}$[q,j] ($j < s_q$) will be reached. Then the time instants from column j to column s_q are the possible end points for the audio segment. Euclidean distance measure have been used to calculate the costs for the matrix $\mathbf{M}$. The above procedure was adapted from a similar approach as mentioned in [4].

The scores corresponding to all the possible end points are considered for *k-means* clustering. For k = 3, mean scores are calculated. Minimum score is used as a threshold to select segments. The segment with the lowest score among the overlapping segments (overlap of 70%) is considered.

2.3 Experimental Results

From the audio content and audio query, speech and non-speech segments are detected. Indexing and DTW search are then applied on the speech segments. Zero filtered signal was generated from the audio signal using a zero frequency resonator. This zero frequency signal is used to detect voiced and unvoiced regions [5]. If duration of unvoiced segment is more than 300ms then it is a non-speech segment.

The system was evaluated based on NIST spoken term detection evaluation scheme [6]. Miss probability and false alarm probability scores on the development data is ranging from 70% - 98% and 0.1% - 0.6% respectively. Miss probability and false alarm probability scores on the evaluation data is ranging from 96% - 98% and 0.1% - 0.2% respectively.

3. DISCUSSIONS

For indexing the audio content, trigrams were used. The approach can be extended for bigram or four-gram. Use of bigram would return a lot of segments and would be a problem due to slow speed of sliding DTW search. In case of four-gram, recall of the number of audio content files has drastically dropped. Trigram seems to strike a balance between bigram and four-gram indexing.

For estimating the end point, k-means clustering was applied on the DTW scores obtained from all the audio segments. We suspect that this might be the reason to loose certain segments. DTW scores obtained from similar kind of pronunciation might mask the scores from the other pronunciation variations.

4. REFERENCES

[1] Arun Kumar, Nitendra Rajput, Dipanjan Chakraborty, Sheetal K. Agarwal, and Amit Anil Nanavati, "WWTW: The World Wide Telecom Web," in *NSDR 2007 (SIGCOMM workshop)*, 2007.

[2] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2006.

[3] Gopalakrishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R.N.V. Sitaram, and S P Kishore, "Development of indian language speech databases for large vocabulary speech recognition systems," in *Proc. SPECOM*, 2005.

[4] Kishore Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," in *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[5] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," in *IEEE Signal Processing Letters*, 2010.

[6] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. of the ACM SIGIR 2007, Workshop in Searching Spontaneous Conversational Speech*, 2007.