College of Engineering
Temple University

Quality Assessment For Search Terms in Spoken Term Detection (STD) Systems

Amir H Harati Nejad Torbati

Department of Electrical and Computer Engineering, Temple University, Philadelphia, Pennsylvania

Abstract

Voice keyword search is an extension of text-based searching that allows users to type keywords and search audio files containing spoken language for their existence.

- Performance dependent on many external factors such as the acoustic channel, language and the confusability of the search term.
- Goal is to develop an algorithm for predicting the reliability or strength of a search term.
- Main objective is to find a predictive function to forecast the quality of given search term.
- Application is to improve the user experience for multimedia information retrieval.
- Assessment is based on the NIST 2006 Spoken Term Detection (STD) evaluation.

STD Approach

- Two general approaches: word-based and phonetics-based
- Goal is to rapidly detect the presence of a *term* in a large audio corpus of heterogeneous speech material
- The effectiveness of a deployed STD system is a tradeoff between processing resource requirements and detection accuracy

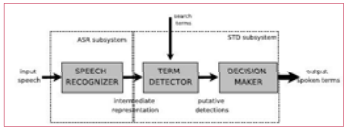
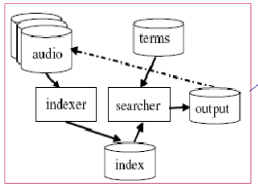


Figure 1. Block Diagram of STD System



A typical STD system divides the task into two separate phases: indexing and searching.

Figure 2. Generic STD System Architecture.

Applications of Spoken Term Detection

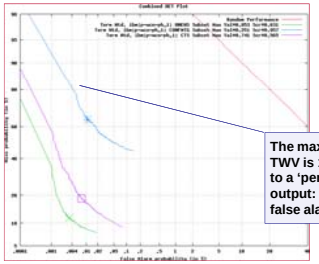
- Simple keyword search
- Complex queries (e.g., logical operators)
- Prioritization of documents to be reviewed
- Spoken document retrieval
- Content-based Internet searches
- Topic Spotting for national security
- Clustering/categorization of spoken documents

Performance Measurement

- Types of Errors: false alarms / missed detections
- Detection error tradeoff (DET) curves are used to evaluate performance.
- A term-weighted value (TWV) criterion was adopted to balance many practical issues..

$$TWV(\theta) = 1 - \underset{terms}{average}\{P_{Miss}(term,\theta) + \beta . P_{FA}(term,\theta)\}$$

$$\beta = \frac{C}{V} . (\Pr_{term}^{-1} - 1)$$



The maximum possible TWV is 1.0, corresponding to a 'perfect' system output: no misses and no false alarms.

Figure 3. Typical STD system performance.

System Block Diagram

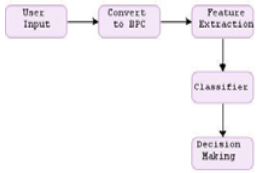


Figure 4. A block diagram of a KWS system.

Methods

- Understanding the relationship between phonetic structure and TWV
- Analyze training set in terms of broad phonetic class and CVC sequences
- Investigate the relationship between different features such as number of syllables, syllable structure and consonant position
- Using linear regression and analysis of variance to select best features
- Using decision trees to understand relationships between features and how to cluster features

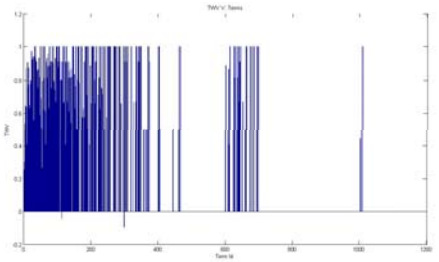


Figure 4. TWV Vs. Terms (from NIST 2006 STD evaluation)

STOPS	b p d t g k
Fricative	jh ch s sh z zh f th v dh hh
Nasals	m n ng en
Liquids	l el r w y
Vowels	iy ih eh ey ae aa aw ay ah ao ax oy ow uh uw er
Gap and Silence	sil sp

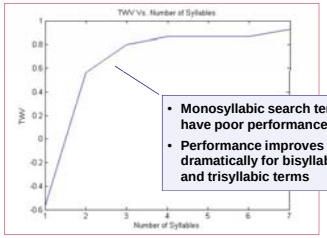
Table 1. Broad Phonetic Class

Initial Phoneme Type	Average TWV
Consonant	-0.25
Vowel	0.44

Table 2. TWV vs. initial phoneme type

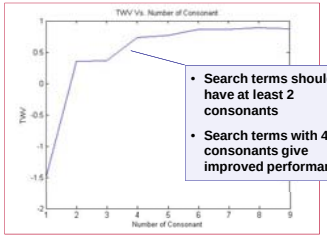
Final Phoneme Type	Average TWV
Consonant	0.43
Vowel	-1.42

Table 3. TWV vs. final phoneme type



- Monosyllabic search terms have poor performance
- Performance improves dramatically for bisyllabic and trisyllabic terms

Figure 5. TWV vs. number of syllables



- Search terms should have at least 2 consonants
- Search terms with 4 consonants give improved performance

Figure 6. TWV vs. number of consonants

Conclusions

- Data-driven approaches to modeling search terms appear promising
- Though the NIST 2006 evaluation data set is small, sufficiently rich generalizations can be made
- Statistical methods such as decision trees will be useful for clustering and evaluating features
- The final prediction function will be based on nonlinear mapping functions trained using machine learning approaches (e.g., neural networks or maximum margin classifiers)

References

1. J. Fiscus, *et al.*, "Results of the 2006 Spoken Term Detection Evaluation," Interspeech 2007.
2. The 2006 Spoken Term Detection Evaluation Plan, http://www.itl.nist.gov/.../std06-evalplan-v10.pdf.
3. D. Wang, *Out of Vocabulary Spoken Term Detection*, PhD Dissertation, Edinburgh University, 2010.
4. Keyword Search Term Prediction, www.isip.piconepress.com/projects/ks_prediction/ (project web site)

Acknowledgments

I am thankful to Sundar Srinivasan and Tao Mao, of Mississippi State University, for their advice and assistance.