Speaker Recognition Using Acoustically Derived Units

Brett R. Wildermoth & Kuldip K. Paliwal

***Abstract*— The field of speaker recognition has been primarily based around text-independent systems based upon GMM. Text-dependent system based on linguistic units have since long been forgotten due to their restrictive nature. It is the goal of this paper to propose the use of acoustic based units as a means of reducing these text-dependent restrictions and hopefully increase the overall system performance.**

I. Introduction

THE topic of speaker recognition encompasses both speaker identification and speaker verification. Speaker identification asks the question *"Who are you?"*, whereas speaker verification asks *"Are you who you claim to be?"*. For both systems speaker modeling and classification is performed in the same fashio., Fowever the application of the classification result differs significantly (see Fig. 1). In a speaker identification system, the speaker is identified through a maximum likelihood search. A speaker is verified by a simple thresholding approach in a speaker verification system.

Speaker recognition systems can be further classified depending on the restrictions enforced on the type of speech that can be spoken to the system. A speaker recognition system with no restrictions is defined as text-independent. Where as a text-dependent system may limit the speaker to a fixed vocabulary or a fixed phrase. These restrictions are a result of the models used by the text-dependent system. These models are generally based upon linguistic units. These types of models consist of whole phrase units, whole word units or sub-word units. Sub-word units consist of linguistic units such as phonemes, di-phones and syllables. These units are related back to the phrase via a lexicon (dictionary) and a word-network.

Instead of using a linguistic based unit with a fixed lexicon it may be more appropriate to use an acoustic based unit with a derived lexicon. In recent years acoustic units have been quite popular in speech recognition [1]–[4], but have never been applied to speaker recognition. It is the goal of this research to explore their usefulness in the speaker recognition context.

Section 2 illustrates how acoustic units can be derived from speech. Section 3 illustrates the performance of speaker recognition systems based on linguistic units. It also cover preliminary experiments undertaken with acoustically segmenting speech. Finally Section 4 concludes the paper.

II. Generating Acoustic Units

The generation of acoustic units is done in three sections. Firstly, the speech is segmented into similar acoustic events. The number of events that the speech is segmented into can be explicitly defined or the number can be increased until a acceptable segmentation error is reached. Once the speech is segmented into M continuous segments, the centroids of each segment are then clustered using a k-means algorithm to reduce the number of acoustic segments to N acoustic units. Using these N acoustic units, label files are generated for the speech and an appropriate lexicon is created for every word in the database's vocabulary.

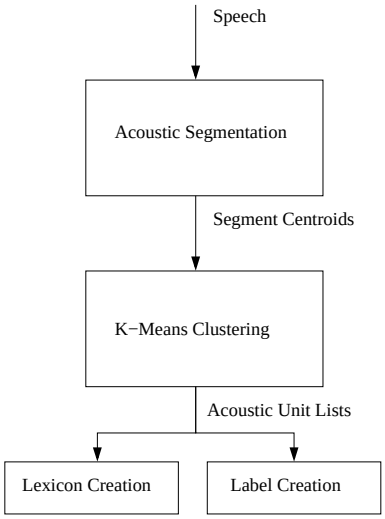

Fig. 2. A simplified look at generating acoustic units.

Acoustic segmentation is performed using dynamic programming where segments are chosen to reduce the overall distortion.

III. Experiments

The performance of speaker recognition system based upon linguistic units is evaluated within the HTK framework. Premilitary results of the acoustic segmentation component of acoustic unit based speaker recognition system are dicussed.

A. Linguistic Units

The performance of three types of linguistic based speaker recognition systems were evaluated. This included a phrase based system, a whole word based system and a phoneme based system.

1) Database: The database used throughout this paper is the Token Evaluation Database (TED). TED is a newly created database for the purpose of comparing linguistic unit based

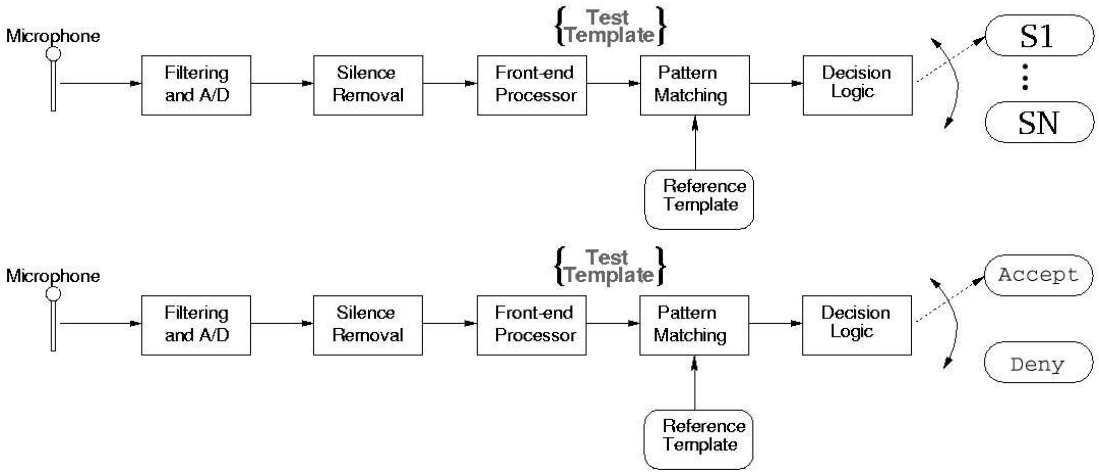

Fig. 1. The Basic Components of a Speaker Recognition (Identification [top], Verification [bottom]) System.

speaker recognition systems against acoustic based systems. The database is still in early stages of development and currently has a population of only six speakers. Once the database has been confirmed as appropriate, the population will be increased. TED consists of a training and testing component.

The training component consists of an utterance of the digits zero to nine repeated ten times. The testing components consists of five sections: Phrase, Word (Post code), Word (Ten Digits), Phoneme, and Free Speech. The phrase component consists of the ten repetitions of the training phrase. The two word components consist of the same words (digits) as in the training phrase, but in a different order. The post code section consists of utterances containing only four digits. The phoneme section consists of ten phrases consisting of ten words containing only the phonemes used in the training phrase. The free speech section consists of five utterances spoken from randomly chosen TIMIT transcripts. Apart from the free speech all participants spoken exactly the same phrases.

The database was recorded at 8kHz with a resolution of 16bits. Label files where automatically generated at word, phoneme and phrase level.

2) Linguistic Unit Based System: Using the speech contained within TED, LPCC features were created over a 20 ms frame and updated every 10 ms. Each frame was pre-emphassied and windowed with a hamming window prior to generating the 12^{th} order LPCCs.

Linguistic modeling and classification was done using HTK version 3.2. Five systems were created within the HTK framework. These included: a fixed phrase speaker recognition system, a prompted and unprompted word based speaker recognition system, and a prompted and unprompted phoneme based speaker recognition system. The prompted systems used a known word network, whereas the unprompted systems had to generate this network via a speech recognition based front-end

3) Experimental Results: The training component of TED was used to generate the linguistic models for each speaker,

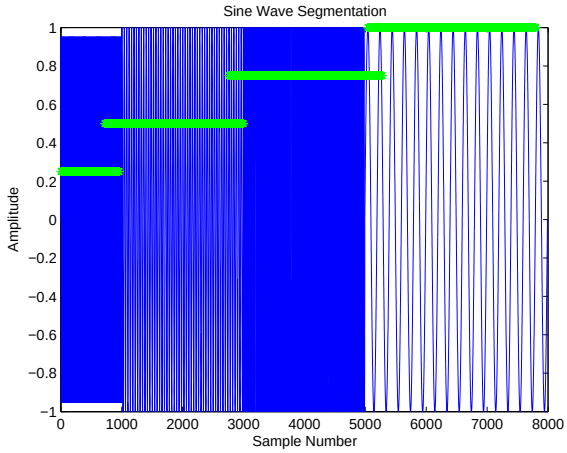

Fig. 3. A sine wave signal segmented into its four frequency components.

and to create general linguistic models for the speech recognition front-end of the unprompted systems. Each component of the TED database was used independently and the results are shown in Table I. From table I it can be seen that all systems functioned extremely well when tested with the same phrase used in training. However, the performance degrades when word or phoneme based speech is used.

B. Acoustic Segmentation

Testing the ability of the system to segment samples into sections of similar acoustic content was tested in three parts. During each test the waveform was reduced to LPCC features [5], generated over a 45 ms frame with an update of 15 ms. Each frame was preemphasised and windowed.

1) Simple Frequency Segmentation: Firstly a waveform consisting of four fixed frequency components was used This was the simplest problem that could be provided and allowed a quick and easy way of confirming the systems functionality. From Fig. 3 it can be seen that the system has

Speech Type	Phrase			Word			Phoneme		
	IER	EER	AR	IER	EER	AR	IER	EER	AR
Fixed Phrase	0.00	0.00	NA	0.00	0.00	NA	0.00	0.00	NA
Fixed Vocab. (Ten digits)	36.67	24.67	2.91	3.34	1.34	2.00	0.00	0.00	NA
Fixed Phoneme Vocab.	68.33	32.00	2.83	55.00	30.50	3.39	53.34	31.67	3.31

TABLE I

PERFORMANCE OF LINGUISTIC BASED SPEAKER RECOGNITION SYSTEMS.

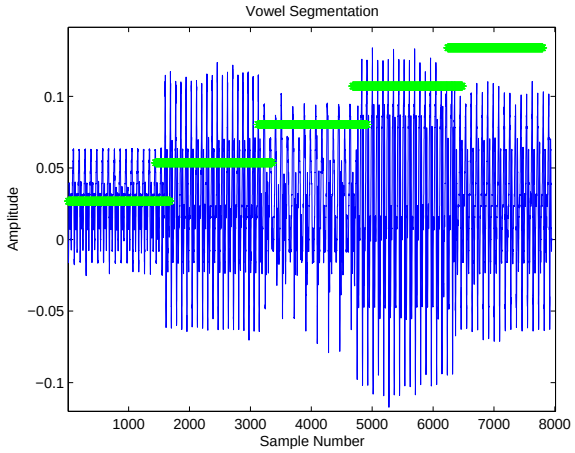

Fig. 4. The vowels sounds a,e,i,o,u segmented.

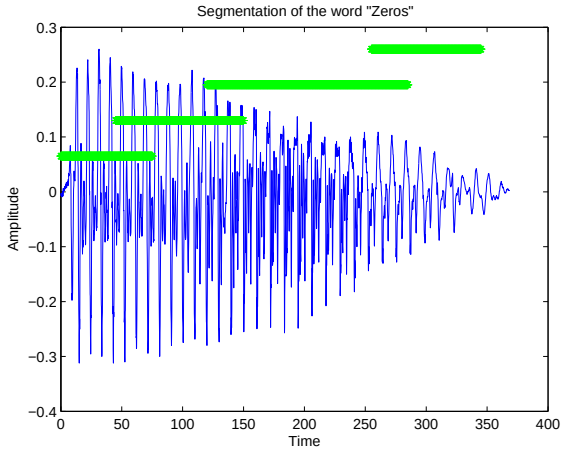

Fig. 5. The digit "Zero" segment into four acoustic segments.

clearly segmented the sine wave into its individual frequency components. The overlap at the boundaries is due to the frame size and frame update used.

2) Simple Speech Segmentation: A sample of speech was artificially created consisting of the five vowels sounds (a,e,i,o,u). The sounds were manually concatenated and all vowels transitions were removed. This sample was then offered to the system. The system was able to correctly segment the speech into its individual vowel components (see Fig. 4).

3) Phoneme Estimation: The next phase of testing the system was to segment a naturally spoken number (namely the digit "zero") into four segments relating to the phonemes *z,i,r,u* contained with in it. This gave an opportunity to compare the acoustic segmentation with the actual phoneme boundaries in this word. The spoken digit came from the training component of the TED database. The digit had the surrounding silence manually removed and LPCC features were generated prior to acoustic segmentation.

The result of the acoustic segmentation can be seen in Fig. 4 and a comparison with the phoneme boundaries is made in Table II. From the table it can be seen that when asked to limit itself to four acoutics events it approximately segments the speech on the phoneme boundaries.

Phoneme	Actual		Segmented	
	Start (ms)	End (ms)	Start (ms)	End (ms)
z	0	90	0	75
i	90	120	45	150
r	120	270	120	285
u	270	345	255	345

TABLE II

COMPARISON OF ACOUSTIC SEGMENTATION TO ACTUAL PHONEME BOUNDARIES.

IV. CONCLUSION

The use of linguistic units enforces certain limitations on the speaker recognition system. Using sub-word units increases the flexibility of the system but at the cost of accuracy and also the need for more training data. The use of acoustic units is yet to be evaluated, but the progress so far is looking promising. The system can successfully segment speech based on its acoustic content. It usefulness in speaker recognition is yet to be evaluated.

REFERENCES

[1] M.A. Bacchiani, *Speech Recognition System Design Based on Automatically Derived Units*, Ph.D. dissertation, 1999, Boston University, 1999.

[2] C.H. Lee, B.H. Juang, F.K. Soong, and L.R. Rabiner, "Word recognition using whole word and subword models," *Proc. ICASSP*, pp. 683–686, 1989.

[3] T. Svendsen, "On the automatic segmentation of speech signals," *Proc. ICASSP*, pp. 77–80, 1987.

[4] K. K. Paliwal, "Lexicon building methods for an acoustic sub-word based speech recognizer," *Proc. ICASSP*, pp. 729–732, 1990.

[5] B.R. Wildermoth and K.K. Paliwal, "Gmm based speaker recognition on readily available databases," *Micro.Elec.Eng. Research Conf.*, 2003.