HARTFEX: A MULTI-DIMENTIONAL SYSTEM OF HMM BASED RECOGNISERS FOR ARTICULATORY FEATURES EXTRACTION

Tarek Abu-Amer and Julie Carson-Berndsen

University College Dublin

ABSTRACT

HARTFEX is a novel system that employs several tiers of HMMs recognisers that work in parallel to extract multi-dimensions of articulatory features. The features segments on the different tiers overlap to account for the co-articulation phenomena. The overlap and precedence relation among features are applied to a phonological parser for further processing. HARTFEX system is built on a modified version of HTK toolkit that allows it to perform multi-thread multi-feature recognition. The system testing results are highly promising. The recognition accuracy for vowel is 98\% and for rhotic is 93%. Current work investigates inherited interdependencies of extracting different feature sets.

1. INTRODUCTION

The Hidden Markov Model based Articulatory Features Extraction (HARTFEX) system uses a multi-dimensional representation of articulatory features. Each dimention represents a feature kind (e.g. manner, place, voicing ...etc). The feature segments on different tiers (dimensions) do not start and end simultaneously and an overlap relations exist among them. This representation is advantageous on that provided by traditional statistical speech recognisers which have been assuming a rigid one dimensional phonemic representation of speech utterances that neglects a lot of the underlined details and does not account for co-articulation phenomena. The multi-dimensional feature representation is fulfilled by means of multiple HMM based recognisers that run in parallel mode for different feature sets. In this stage of our work the HMM recognisers are mutually independent and no information are exchanged among them. The HMM recognisers rely on multi-thread paradigms that take care of synchronisation and resources sharing. HARTFEX system is designed to be the first stage of a computational linguistic model for speech recognition, Time Map Model (depicted in figure 1), where the processing goes from the feature extraction stage to the phonological parser stage and finally to a lexicon search stage.

2. THE MULTI-DIMENTIONAL REPRESENTATION OF SPEECH

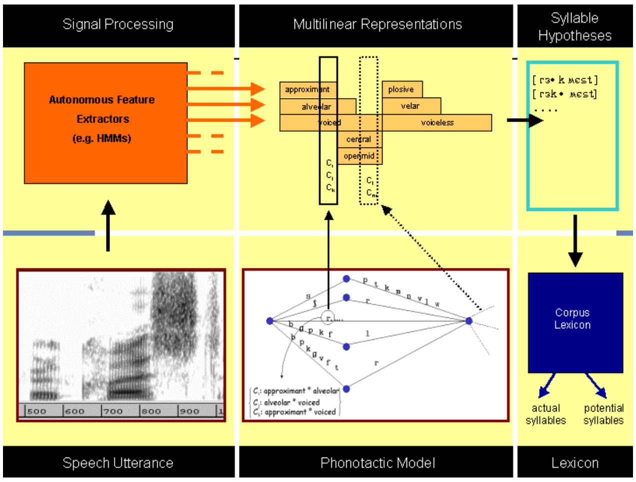

Figure 1 TIME MAP Model

Speech utterances in HARTFEX system are defined in terms of a multi-dimensional representation of articulatory features. Each diminution represents a feature kind that is associated with signal time. This representation enriches the information gathered from the speech signal instead of just gathering a poor one-dimensional phonemic representation. The features on different dimensions do not all start and end simultaneously. An overlap between features on different dimensions represents an event that is further processed by the phonological parser. for example, in figure 2 the feature {**rd-**} begins before the feature {**voc**} indicating that the lips have been spread during the plosive {**stp**} anticipating the following vowel. It can be seen how this representation of features captures coarticulation phenomena.

3. HARTFEX SYSTEM OVERVIEW

The HARTFEX system uses a modified version of Cambridge HTK toolkit that allows several HMMs based recognisers to run in parallel executing threads to extract different features dimensions. Six feature dimensions are defined in HARTFEX system, namely: manner of articulation, place of articulation, voicing, vowel type, vowel height and lip round. The HMMs design topology used is as follows: 1) Context independent HMMs were built to model individual features. 2) an HMM is reserved to model silence on each feature dimension. 3) the number of states per HMM is **5** for all feature dimensions. 4) multiple

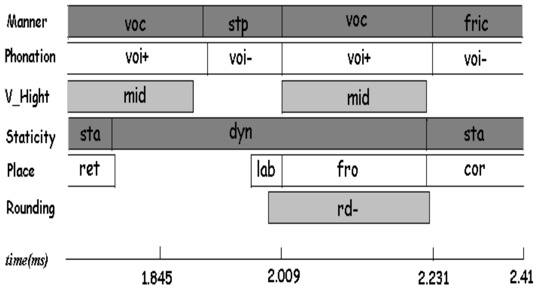

Figure 2 Multi-dimensional representation of *pace*

Gaussian mixtures are used to model output distribution in each state on an HMM. The Manner of articulation feature classes needed **5** mixtures per state for best performance. Other feature dimensions needed **15** mixtures per state for best performance. The six HMMs system were trained off-line using a corpus of American English (**TIMIT**). *Baum-Welch* (forward-backward) algorithm was used for training. The number of training iterations after each change in the HMMs was restricted to two iterations in order not to fall into over fitting to the training data. The HARTFEX system works for both live and batch mode recognition. Viterbi algorithm was used for searching.

4. SPEECH SIGNAL PROCESSING IN HARTFEX SYSTEM

It is will discovered that the human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggest that designing a front-end to operate in a similar non-linear manner improves recognition performance. A filter bank analysis provides a much more straightforward route to obtaining the desired non-linear frequency resolution. However filter bank amplitudes are highly correlated and hence, the use of a cepstral transformation in this case is virtually mandatory if the data is to be used in a HMM based recogniser with diagonal covariances. In HARTFEX system *Mel frequency cepstral coefficients* (**MFCC**) technique was used for parameterising speech frames as it was found to give a good discrimination and so deliver the best performance. Speech frames are captured using a 25ms sliding window with 10ms overlap between successive windows. The observation vector is comprised of 12 MFCC static coefficients,12 deltas, 12 accelerations and 3 energy components.

5. RECONSTRUCTING HTK TOOLKIT FOR HARTFEX SYSTEM

HTK *(Cambridge Hidden Markov Model Toolkit*) is comprised of tools for speech analysis, HMM training, testing and result analysis. Much of the functionality of HTK is built into the library modules that provide a central resource of commonly used function. HTK library modules were reconstructed to accommodate a multi thread environment. **<u>HMem</u>** the memory management library, needed specifically to be manipulated in order to have dedicated memory heaps for each thread and make memory allocation/de-allocation thread safe.**Hvite**, the **HTK** recognition tool was reconstructed to supply multiple concurrent recognisers for the different feature dimensions. The new multi-thread *Hvite* (*THVite*) has three types of executing threads as follows: **<u>The Main Thread</u>**: which deals with the shell, creates other working threads ,wait until they finish and then exits. **<u>The Recording Thread</u>**: which interfaces to the audio source, performs automatic speech/silence detection, performs parameterisation and concatenates all observations belonging to one utterance in one large buffer (this is a new modification to HTK) before passing it to the recognising threads. **<u>Recognising Threads</u>**: that receive control from the Main thread ,initialises, read observations from the recording thread, perform Viterbi search, output utterance recognition results concurrently and then bind for the next input.

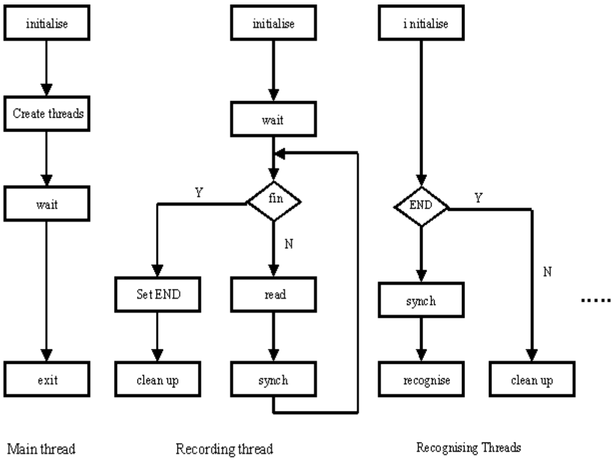

Figure 3 *Thvite* threads

6. HARTFEX SYSTEM TRAINING AND TESTING

The HARTFEX system was trained and tested on continuous speech utterances from TIMIT corpus. The TIMIT phonemic transcription was transformed into the articulatory features transcription using the Generic Transducer Interpreter (an in-house tool that uses phonotactic constraints to translate input symbols into output ones). The TIMIT training set is composed of 4620 utterances spoken by 462 speakers. The testing set is composed of 1344 utterances spoken by 168 speakers. No utterances appeared both in training and testing. The system performance is extremely promising and is advantageous on other recent approaches for articulatory feature extraction. So far the

tier name	feature classes recognition results								
place	sil	vel	fro	cen	cor	bak	ret	lab	den
	99	89	92	75	80	80	93	83	84
manner	sil	voc	nas	stp	frc	flp			
	100	98	90	89	82	88			
vowel height	nil	Hi	mid	lo					
	97	89	76	92					
vowel type	nil	Lax	ten						
	92	92	88						
round	nil	rd-	rd+						
	97	92	88						
voice	vo+	vo-							
	90	94							

Tables 1 HARTFIX testing results for : (1) manner, (2) place, (3) voicing, (4) vowel type, (5)vowel height and (6) rounding (*nil* on the vowel related tiers models non-vowel segments)

recognisers working for the different feature dimension work completely independent on one another. Despite this fact the recognition accuracy is noticeably high. The testing results for recognising the different feature classes are reported in table 1. It is evident from the manner results in table 1 and also from the confusion matrix of the manner in table 2 that recognition accuracy for vowel segments is very high(98%). This suggest dividing the extraction process to two steps: first extracting manner features to distinguish vowel segments from consonant segments. Then using this knowledge to increase the accuracy of extracting place features and in the same step introducing only vowel segments to vowel related recognisers(vowel type, vowel height and lip round).

7. CONCLUSION AND FUTURE WORK

The HARTFEX system is a novel approach to extract multi-dimensional articulatory features by means of multi-layered systems of HMMs. It is built upon a novel multi-thread version of HTK toolkit. The system has multi-dimensional and non-linear approaches to speech recognition inherited in it. The system delivers very promising recognition accuracy while the different features classes recognisers are independent on one another. The system capabilities can be further extended by exploring the possible inter-dependencies of extracting the different feature classes. Also the performance can be refined by further considering the non-linear properties inherited in the speech signal. A specific issue in the respect is the treatment of speech signal as chaos and the possibility of using HMMs to model chaos signals. Another attractive aspect about HARTFEX is that it relies on articulatory features that are common to most languages of the world which makes it inherently cross-linguistic in capability and extendible to other corpora.

	Sil	**voc**	**nas**	**stp**	**frc**	**flp**
sil	**2686**	0	0	0	0	0
voc	0	**13444**	131	44	18	39
nas	0	74	**3476**	55	0	245
stp	0	74	138	**5424**	140	335
frc	0	107	150	558	**5560**	445
flp	0	9	47	23	20	**751**

Table 2 A confusion matrix illustrating The classification performance of manner of articulation recogniser (correct hits, are marked in **bold).**

8. REFERENCES

[1] Carson-Berndsen, J., Time Map Phonology: Finite State Models and Event Logics in speech Recognition, Kluwer Academic Publisher, Dordrecht, 1998.

[2] Carson-Berndsen, J., "Finite State Models, Event Logics and Statistics in Speech Recognition", In: Gazdar, G.; K. Sparck Jones & R. Needham (eds.):Computer, Language and Speech: Integrating formal theories and statistical data. Philosophical Transactions of the Royal society, Series A, 358(1770), 1255-1226.

[3] Steve Young, Phil Woodland and Gunnar Evermann, HTK Book, Cambridge University Engineering Department 2002.

[4] Steve Young and Gerrit Bloothooft, Corpus-Based Methods In Language And Speech Engineering, Kluwer Academic Publisher, Dordrecht, 1997.

[5] Chang, S;S. Greenberg & M. Wester, "AnElitist Approach to Articulatory-Acoustic Feature Classification", In: Proceedings of Eurospeech 2001, Aalborg.

[6] Ali, A.M..A, J. Van der Spiegel, P. Mueller, G Haentjaents & J. Berman, "An Acoustic-Phonetic Feature-

[7] Kai-Fu Lee and Raj Reddy, Automatic Speech Recognition: The Development of the Sphinx Recognition System, Kluwer International Series in Engineering and computer Science II , Dordrecht,1989.