A MULTI-LINEAR HMMs SYSTEM FOR ARTICULATORY FEATURES EXTRACTION

Tarek Abu-Amer and Julie Carson-Berndsen

University College Dublin

ABSTRACT

In this paper we present a novel approach for articulatory features extraction. Our approach relies on an auto-segmental multilinear representation of AFs. Overlap and precedence relations between AFs on the different levels of the multi-linear representation can be extracted and then presented to a phonological parser for further recognition. This representation models co-articulation affect. We used parallel systems of multi-Gaussian Hidden Markov Model based recognisers to extract AFs classes. The statistical system was implemented using a novel modification of the HTK toolkit which allows it to perform multi-thread multi-feature recognition. Testing proved the overall performance is extremely promising. Among the highest accuracies achieved are 98% for vowels and 93% for rhotic sounds. Current work investigates interdependencies of extracting different feature types.

1. INTRODUCTION

The term Articulatory Features subsumes a variety of concepts, ranging from features which are typically used in linguistic phonological systems to categorise speech sounds to acoustic properties found in speech signal[8]. It is better to use articulatory features for ASR rather than using phones for the following reasons: 1) AF bear relation to the speech signal as well as to higher level linguistic units. 2) Individual AF exhibit much less variations than phones so it would be easier to model them and so to recognise them. 3) They permit more efficient exploitation of available training data. Therefore the statistical feature models can be trained more robustly. 4) AF can successfully model co-articulation affects. In spite of their advantages, AFs were rarely used in statistical ASR systems. Typical commercial ASR systems have proven successful but they have usually used traditional phonemic representation that is segmentation based and so can hardly model co-articulation affect. In Our approach we take the advantages of statistical modelling and of AFs in a representation that outperforms the typical ASR systems and essentially is: Auto-segmental, Multi-linear and also makes use of the available phonemically transcribed databases like TIMIT. In the following we give an overview of our system followed by a detailed explanation of the technology beneath it.

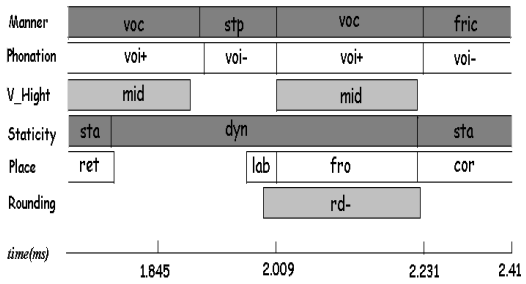

Figure 1 Multi-linear representation of *pace.*

2. SYSTEM OVERVIEW

HARTFEX: The **H**MMs based **ART**iculatory **F**eatures **EX**tractoion system defines an auto-segmental multi-linear representation of AFs. As can be shown from figure 1, each feature in a multi-linear feature representation is associated with a specific tier (on the vertical axis) and with specific time interval in terms of milliseconds (on the horizontal axis). The features do not all start and end simultaneously. An overlap of properties exists in any time interval; for example, in figure 1 the feature **rd-** begins before the **voc** feature indicating that the lips have been spread during the plosive **stp** anticipating the following **vowel**. In this way, the multi-linear feature representation models co-articulation phenomena. This representation assumes a non-segmental approach to the description and interpretation of speech utterances which avoids having to segment an utterance into non-overlapping units at any level of representation. Six tiers of features are defined in HARTFEX: 1) The Manner of articulation tier ,2) The Place of articulation tier, 3) the voicing tier, 4) the vowel type tier, 5) the vowel height tier and 6) the round tier. The AFs are modelled by parallel systems of HMMs. The HMMs systems are independent of one another and work on separate execution threads. However, they share the same system resources so specific multi-thread synchronisation paradigms are needed to guide the processing of the recognition threads. HARTFEX system forms one component of a computational linguistic model for speech recognition, the *Time Map Model* (depicted in figure 2), which was first proposed by Carson-Berndsen [1][2]. The *Time Map Model uses* a finite state network representation of

ISBN 1-876346-48-5

the phonotactic constraints in a language, known as a phonotactic automaton, together with event logic to interpret multilinear representations of speech utterances. The phonological parser of the *Time Map Model* interprets the output of HARTFEX (the articulatory feature tiers) distinguishing between well-formed or ill-formed syllables. The HARTFEX system explores the power of HMMs and defines a generic multilinear representation of speech. It represents a hybrid approach that gets both training-based and rule-based methods of speech recognition to work together in one composite system. The HARTFEX system has thus far been trained and tested using the TIMIT corpus of American English sentences.

3. THE STATISITCAL APPROACH IN HARTFEX

The HARTFEX system extends the Cambridge ***HTK*** toolkit [3][4] by reconstructing the modules so that they can accommodate several HMMs based recognisers that run in parallel to extract different articulatory feature sets. Six HMM systems were built to model the features of the following tiers : 1) manner, 2) place, 3) voicing, 4) vowel type, 5) vowel height and 6) lip rounding. The individual feature classes on any tier are modelled by context independent HMMs. An HMM is dedicated to model the silence in each feature tier. The number of states inside any HMM was set to 5. The output distributions inside each state are modelled by multiple mixture Gaussian distributions. During system training the number of mixtures inside each state is gradually increased and the recognition performance is measured until it saturates at a certain number of mixtures. The manner tier recogniser saturates at 5 mixtures per state, other tiers recogniser need 15 mixtures per state for their performance to saturate. The speech signal is captured using 25ms sliding window with 10ms overlap between successive windows. Mel frequency cepstral coefficients (MFCC) technique was chosen for parameterising speech windows as it achieves good discrimination and so delivers the best performance. The observation vector is comprised of 12 MFCC static coefficients, 12 deltas, 12 accelerations and 3 energy components. The 0^{th} cepstral component is discarded in case of voicing feature recognition. This component represents the vocal tract response (the slow varying signal) while voicing describes the excitation coming from the vocal cords alone (the fast varying signal). The speech signal is pre-emphasised to deal with the attenuation that takes place when speech is emitted from the lips. HMMs off-line training was done using *Baum-Welch* (forward- backward) algorithm. This process attempts to find a maximum likelihood estimate (MLE) for the model parameters; i.e. a parameter set maximising the probability of the training data The number of training iterations after each change in the HMMs was restricted to two iterations in order not to fall

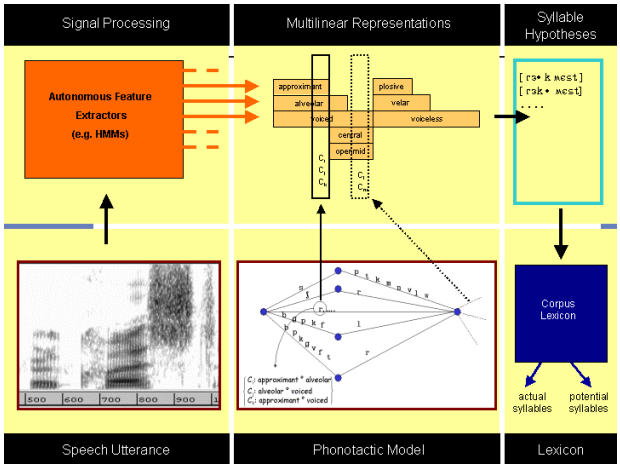

Figure 2 TIME MAP Model

into over-fitting to the training data. The HARTFEX system is designed to work both for live and batch mode speech recognition. The algorithm used for searching during the recognition process is Viterbi algorithm which works to find the most likely state sequence through any model for the given observation vectors.

4. *t*HTK: THE MULTI-THREAD HTK TOOLKIT

The Hidden Markov Model Toolkit (HTK) is the Cambridge toolkit for building and manipulating hidden Markov models [3][4]. HTK is composed of tools that provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. HTK contains a number of tools that do all the work related to building and testing HMM systems. Much of the functionality of HTK is built into the library modules. These modules ensure that every tool interfaces to the outside world in exactly the same way. They also provide a central resource of commonly used functions. To suit the HARTFEX multiple recogniser's paradigm some library modules needed to be altered to suite a multi-thread environment. ***HMem*** 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. The HTK tools used for training HMMs remained unchanged as HMMs systems for different feature kinds are trained individually. The HTK recognition tool ***Hvite*** was reconstructed to supply multiple concurrent recognisers for the different feature sets. The new multi-thread version of ***Hvite*** is ***Thvite***. It was organised into a *main thread* , a *recording thread* and a number of *recognising threads*. The main thread deals with shell, create other working threads and wait until they finish then it exits. The recording thread interfaces to the audio

phone	manner	place	voice	vowel	height	round
[b]	stop	lab	Voi+	-	-	-
[d]	stop	cor	voi+	-	-	-
[g]	stop	vel	voi+	-	-	-
[p]	stop	lab	voi-	-	-	-
[t]	stop	cor	voi-	-	-	-
[k]	stop	vel	voi-	-	-	-
[dx]	flap	cor	voi+	-	-	-
[jh]	fric	cor	voi+	-	-	-
[ch]	fric	cor	voi-	-	-	-
[s]	fric	cor	voi-	-	-	-
[sh]	fric	vel	voi-	-	-	-
[z]	fric	cor	voi+	-	-	-
[zh]	fric	vel	voi+	-	-	-
[f]	fric	lab	voi-	-	-	-
[th]	fric	den	voi-	-	-	-
[v]	fric	lab	voi+	-	-	-
[dh]	fric	den	voi+	-	-	-
[m]	nas	lab	voi+	-	-	-
[em]	nas	lab	voi+	-	-	-
[n]	nas	cor	voi+	-	-	-
[nx]	flap	cor	voi+	-	-	-
[ng]	nas	vel	voi+	-	-	-
[en]	nas	cor	voi+	-	-	-
[l]	v_a	cen	voi+	ten	mid	-
[el]	v_a	cen	voi+	lax	mid	-
[r]	v_a	ret	voi+	ten	mid	-
[w]	v_a	bak	voi+	ten	hi	rd+
[y]	v_a	fro	voi+	ten	hi	rd-
[hh]	fric	glo	voi-	-	-	-
[hv]	v_a	cen	voi+	lax	mid	-
[iy]	v_a	fro	voi+	ten	hi	rd-
[ih]	v_a	fro	voi+	lax	hi	rd-
[eh]	v_a	fro	voi+	lax	mid	rd-
[ey]	v_a	fro	voi+	ten	mid	rd-
[ae]	v_a	fro	voi+	ten	lo	rd-
[aa]	v_a	cen	voi+	ten	lo	rd-
[aw]	v_a	cen	voi+	ten	lo	rd+
[ay]	v_a	cen	voi+	ten	lo	rd-
[ah]	v_a	cen	voi+	ten	lo	rd-
[ao]	v_a	bak	voi+	ten	lo	rd-
[oy]	v_a	bak	voi+	ten	mid	rd-
[ow]	v_a	bak	voi+	ten	mid	rd+
[uh]	v_a	bak	voi+	lax	hi	rd-
[uw]	v_a	bak	voi+	ten	hi	rd+
[er]	v_a	ret	voi+	lax	mid	-
[axr]	v_a	ret	voi+	lax	mid	-
[ax]	v_a	cen	voi+	lax	mid	rd-
[ix]	v_a	fro	voi+	lax	hi	rd-
[q]	stop	glo	voi-	-	-	-

Table1 Mapping of TIMIT phonemes into the articulatory features used for training and testing of the HARTFEX system.

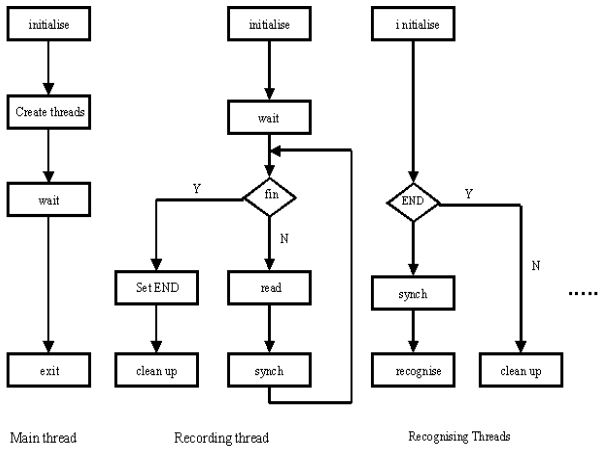

Figure 3 *Thvite* threads

source, performs automatic speech/silence detection, performs parameterisation and concatenates all observations belonging to one utterance in one large observation buffer (which is a new modification to HTK) before passing it to the recognising threads. The recognising threads receive control parameters from Main Thread, initialise threads decoding modules concurrently, read observations from Recording Thread, perform Viterbi search and Output utterance recognition result concurrently and then bind for the next input. The interactions of the Thvite threads are depicted in fig. 4. The multi-threading paradigms uses POSIX Threads. A synchronisation gate was implemented so as recognising threads wait for each other after performing recognition concurrently before triggering the recording thread to fetch the new utterance.

5. HARTFEX TRAINING AND TESTING

The HARTFEX system was trained and tested on continuous speech utterances from the TIMIT database. The TIMIT phonemic transcription was transformed into the arcitulatory features transcription using the *Generic Transducer Interpreter* (an in-house tool that uses phonotactic constraints to translate input symbols into output ones). Table 1. gives the mapping from TIMIT phones into the articulatory features used for the HARTFEX system training and testing. TIMIT recordings cover 8 major dialect regions in the US. Only the phonemically-compact and phonemically-diverse sentences were used for training and testing HARTFIX system. The 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.

ISBN 1-876346-48-5

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 2 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)

6. TESTING RESULTS

The testing results for recognising the different feature classes are reported in table 2. The system performance is extremely promising and outperforms other recent approaches for articulatory features extraction [5][6]. The recognition accuracy is high given that so far the feature sets' recognisers are completely independent on one another as there is no knowledge exchanged from any feature set recognition that can further aid the recognition of other feature sets. It can be seen from the manner tier results in table 2 and from the confusion matrix of the manner tier in table 3, the system's very high accuracy of recognising vowel segments (98%). This encourages the use of this knowledge to extract features on other tiers. For example only the vowel segments would be presented to vowel related tiers'(vowel type and vowel height) recognisers. This way we can get rid of the (*nil*) models representing non-vowel segments when extracting vowel-related features

7. CONCLUSION AND FUTURE WORK

The HARTFEX system is a novel approach to extract articulatory features by multi-layered systems of HMMs. It defines a generic auto-segmental multi-tiered feature representation. It is built upon a novel multi-threaded version of HTK toolkit. The system delivers very promising recognition accuracy while the recognisers for different feature tiers work independently of one another. The system capabilities can possibly be further extended by exploring the possible inter-dependencies for extracting different features tiers. One paradigm [5] suggests extracting manner classes first to distinguish between vowels and consonants segments then extract other features classes giving their manner of articulation. An 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 3 A confusion matrix illustrating the classification performance of manner of articulation recogniser (correct hits, are marked in **bold).**

9. 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.

[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, "An Elitist 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-Based System for Automatic Phoneme Recognition in Continuous Speech", In: IEEE International Symposium on Circuits and Systems (ISCAS-99), III-118 – III-121, 1999.

[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.

[8] Kirchhoff K, Integrating Articulatory Features Into Acoustic Models For Speech Recognition, Univ. of Washington, Seatelle.

10. ACKNOWLEDGMENT

This work was funded by Enterprise Ireland under Grant No.IF/2001/021.