Comparative Analysis of a Continuous and Discontinuous Piecewise Linear Decoder

Chloe Seivwright, Martin Russell, Steve Houghton
School of Electronic Electrical and Systems Engineering, University of Birmingham

UNIVERSITY OF BIRMINGHAM

Aims

The aim of this work is to incorporate a more faithful model of speech for recognition. This recognition task is considered where small amounts of data is accessible for training and developing the system. Therefore, it is important that the model design maintains a small number of parameters.

These models address the independence assumption that is inherent in conventional HMM recognition methods. Motivated by the works of Weber et. al. (2014) [4] a continuous single state HMM (CS-HMM) has been implemented in this work.

CS-HMM Detail

A continuous state HMM uses the Holmes-Mattingley-Shearme (HMS) model used for speech synthesis [3], and applies this method to the recognition problem.
It uses a piecewise linear approximation of the data producing smooth continuous trajectories, representing the smooth and continuous movement of articulators during speech.

A state in this model (x) is compromised of both discrete and continuous components. Each hypothesis contains information on a mean and precision that characterises a Gaussian distribution for that particular hypothesis. As data is observed, the most probable hypotheses are updated to contain all known information up to and including the current observation.

A compact iterative decoding algorithm described in full in [2] is currently implemented in the two systems. This algorithm is a forward alpha pass, updating the parameters of the distribution and also keeping a 'score' (K_t) which is like the sum of the probabilities:

$$\alpha_t(x) = K_t N(x - \mu_t, P_t)$$

PL & PLC Models

There are two systems that are being compared. A disconnected piecewise linear model (PL), and connected piecewise linear model (PLC). The difference being a continuity constraint is enforced at the segment boundaries in the PLC system.

- PLC system is continuous throughout and is a more faithful model of speech, therefore in theory is expected to give the best performance.
- On the full TIMIT 'test' data, **PL** system performed better with an accuracy of **53.54**% compared to **PLC** – **50.97**%.
- Evaluate the difference in performance between the two systems, can any conclusions be drawn from the results?
- Graphically analysing the output of the two systems.
- A statistical binomial significance test has been done to confirm the trend in the data.

Input Data

A low dimensional representation of the TIMIT data is extracted from the bottleneck of a 5 layer Neural Network. See [1] for detail. The system is trained and tested using this data.

Experiment Results

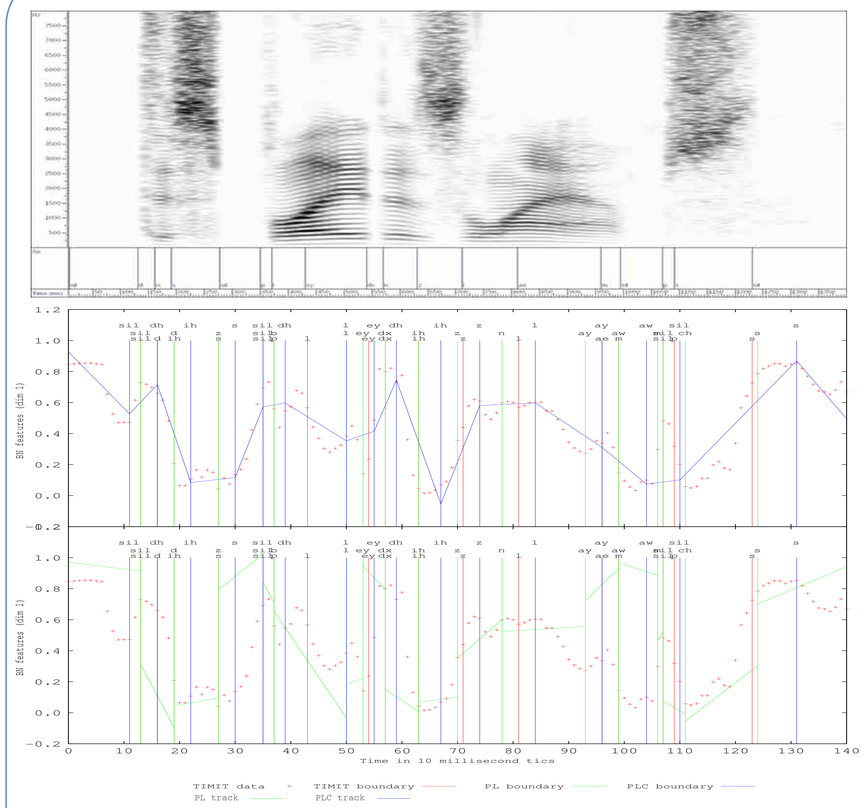

Figure 1: (a) Narrow band spectrum of 1.4 seconds of file `TIMIT/TEST/DR7/MERSO/SI497`. (b) PLC and PL Decoder output recovery tracks with TIMIT, PL and PLC timing boundaries.

Figure 1 show the first 1.4 seconds of `TIMIT/TEST/DR7/MERSO/SI497` file. The recognition results of this experiment are:

System Summary Results						
Data	N	H	Sub	Del	Ins	Acc (%)
PL	69	41	16	12	4	53.62
PLC	69	40	15	14	6	49.28

By looking at the output of the two systems Fig 1(b) we can see that for this first dimension of the feature data, the tracks fit the data relatively well.

The PL system matches the TIMIT boundaries more than the PLC system who's boundaries seem to be shifted throughout the utterance.

The spectrogram (Fig 1(a)) highlights regions of discontinuities in the utterance, the PL system copes with these discontinuities better than the PLC by segmenting the data as a compromise between the TIMIT and PLC boundaries.

Significance Test Results

A statistical binomial significance test using the full recognition confusion matrices as input and identifying which confusions were significantly greater or less in one system or the other. The results shown in the following table show the % correct achieved per phoneme for the PL and PLC systems on the Full TIMIT 'test' set.

Phonem	PL %c	PLC %c	Phoneme	PL %c	PLC %c	Phoneme	PL %c	PLC %c
aa	64.1	68	f	77.6	78.4	oy	68.6	73.7
ae	54.1	58.7	g	59	45.8	p	75.6	73.9
ah	48.5	50.6	hh	71.6	74.3	r	62.2	58.4
aw	57.5	69.3	ih	40.2	38.3	s	83.8	90.8
ay	69.5	78.7	ix	50.4	51.8	sh	81.3	83.9
b	77.8	42.3	iy	74.2	80.3	sil	88.3	83
ch	82.3	88.4	jh	61.5	53.3	t	71.5	68.9
d	50.9	17.6	k	77.2	73.7	th	49.4	47
dh	39.8	62.7	l	62	68	uh	30.4	19.9
dx	78.7	66.6	m	69.5	72.3	uw	56.5	60.2
eh	48.1	48.5	n	67	56.4	v	57.4	36.8
er	67.4	74.2	ng	77.5	71.2	w	69.2	75.3
ey	70.1	79.2	ow	44	49.2	y	62	50.5

Conclusions

A number of conclusions can be drawn from these experiments:

- Despite speech being a continuous process, the discontinuous PL system performs better than the PLC system.
- The PL system provides a better approximation of the data in regions of consonant bursts and closures whereas the PLC system performs better in voiced regions.
- A consequence of enforcing continuity is that the segment boundaries are shifted from the TIMIT segmentation.
- The spectrogram highlights abrupt changes in energies between consonants, the bottleneck data in these regions also show this discontinuity hence the PL system working better when compared to the PLC.
- A hybrid of the PL and PLC systems has potential to be an optimal decoder where some discontinuity is expected, the significance test results identify these regions.
- Future work will involve implementing these systems using three state models where the recognition result is expected to significantly improve.

References

1 Linxue Bai, Peter Jančovič, Martin J Russell and Philip Weber. "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics". *[Accepted INTERSPEECH* 2015].

2 Colin Champion and Steve Houghton. "Application of continuous state Hidden Markov Models to a classical problem in speech recognition". Computer Speech & Language, 2015

3 John N Holmes, Ignatius G Mattingly, and John N Shearme. "Speech synthesis by rule". Language and speech, 7(3): 127–143, 1964.

4 Philip Weber, Steve Houghton, Colin Champion, Martin Russell, and Peter Jančovič. "Trajectory analysis of speech using continuous state Hidden Markov Models". ICASSP 2014 - Speech and Language Processing (ICASSP2014 - SLTC), Florence, Italy, 2014.

Chloe Seivwright
CAS390@bham.ac.uk