COMPARATIVE PERFORMANCE ANALYSIS OF STATISTICAL TRAJECTORY MODELS IN CELLULAR ENVIRONMENT

Bojan Petek [*], *Ove Andersen, and Paul Dalsgaard*

Center for PersonKommunikation, Aalborg University, Denmark

ABSTRACT

Two systems (Statistical Trajectory Models (STM) and continuous density HMMs) utilizing three preprocessing methodologies (MFCC, RASTA and FBDYN) were evaluated on two databases, namely CTIMIT and the corresponding downsampled TIMIT. Within the bounds of the experimental setup the comparative performance analysis showed that the STM significantly outperforms the HMM system on the CTIMIT database. Specifically, the performance of the STM system was found to be at least 10% better as compared to the one obtained by HMM when the RASTA preprocessing was used. The performance of both systems with FBDYN parametrization was found to be inferior to those using MFCC and RASTA. On the other hand, in low-noise conditions on the TIMIT database FBDYN yielded an improved performance for the HMM system, whereas STM achieved the best results with the MFCC parametrization.

1. INTRODUCTION

Statistical Trajectory Models (STM) have been shown to be a promising approach in automatic speech recognition at low noise-level conditions [4]. To date, however, no research has been reported which evaluates their performance in celluar environment with its inherent low signal-to-noise ratios and occasionally drop outs.

Due to the increasing global importance of mobile communications and the availability of a multi-speaker, continuous speech database collected over the cellular network CTIMIT [2], it is clear that the robust speech recognition and processing in cellular environments will become an important and challenging subject for research.

Based on these facts we have decided to evaluate the effectiveness of an alternative approach of STM and the traditional HMMs in a comparative performance analysis. Our experiments encompasses two databases (CTIMIT and the corresponding downsampled (8kHz) TIMIT), two systems (STM [4] and the HTK v2.0 [9]), and three preprocessing methodologies (MFCC, RASTA [5], and FBDYN [1]).

[*] Visiting researcher from the University of Ljubljana, Slovenia and postdoctoral scholar of the Slovenian Science Foundation.

The paper is organized as follows. Section 2 highlights the major characteristics of the CTIMIT and the corresponding downsampled TIMIT databases. Section 3 briefly overviews the STM and HMM systems chosen for the experimental part of this work. Results from the comparative performance analysis are divided into two parts and summarized in Section 4. The first part (Section 4.1) discusses the overall performace scores in phonetic classification experiments. The second part (Section 4.2) summarizes results from the rank-order statistics on the confusion matrices for each of the experiments. Finally, Section 5 presents conclusions and some possible future research directions.

2. DATABASES

CTIMIT corpus [2] is a cellular bandwidth complement to the TIMIT database [7]. Its creation is in principle similar to that of the NTIMIT (*n*etwork TIMIT [6]), i.e. the TIMIT, originally recorded under clean channel conditions, was transmitted over the cellular network, digitized at a rate of 8 kHz, and organized into a directory structure corresponding to that of the TIMIT database. Thus, CTIMIT maintains the carefully designed phonetic coverage coupled with the effects added by cellular communication environments and transmission characteristics.

We have used the 1.0 alpha version of CTIMIT which consists of 3367 (out of 6300) TIMIT sentences. Other limitations in using this corpus include the use of a single vehicle in collection, a single receiving phone line, a limited number of cell phones used, no hands-free mode simulation, and a lack of documentation of call conditions. Online documentation, however, specifies that the last two (out of 7) recording sessions are relatively low quality cellular channels characterized by considerable interference and a high dropout rate. An example of this is depicted in Figure 1. Two main conclusions can be drawn from this. Firstly, if such an utterance is provided to the system analyzed in this paper as a training token, the corresponding phonetic transcription needs to be *respecified* in order to maintain a consistent model training. Secondly, dropouts need to be handeled in the recognition framework to achieve robustness

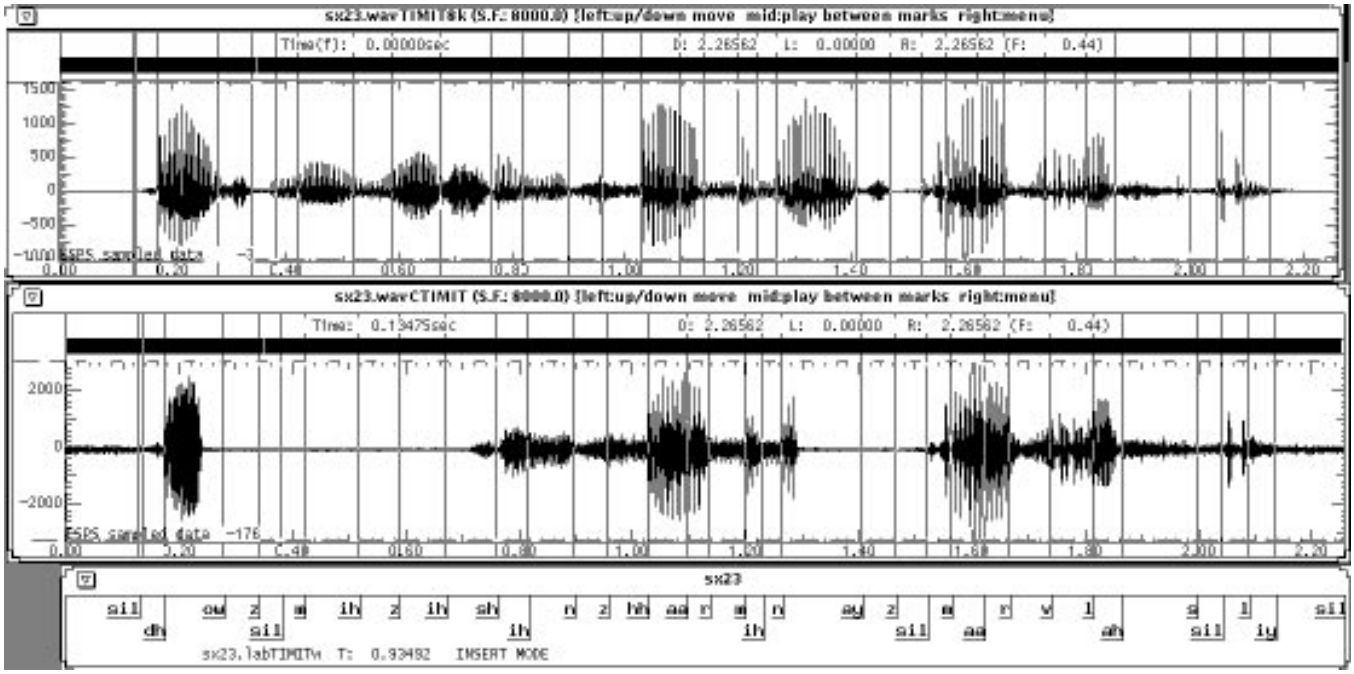

Figure 1: An example of a dropout present in the CTIMIT database. Upper trace: Original TIMIT utterance. Middle trace: corresponding CTIMIT utterance. Lower trace: phonetic labels. The example shows that complete phoneme sequences are missing or portions of phonemes are cut in the CTIMIT utterance.

due to randomly missing data.

In order to make the analysis as realistic as possible, we downsampled the 3367 original TIMIT utterances to an 8 kHz sampling rate to match that of the CTIMIT. The delay of a downsampling FIR filter was compensated for in order to preserve proper alignment of the phonetic transcriptions. 2465 utterances present in the training part of the database were used as a training set, and another 902 in the test part as a test set.

3. SYSTEMS

Recently, several alternative modeling paradigms, broadly classified as *segment models* [8], have been proposed within the area of ASR. Segment models may be considered as generalized HMMs, i.e. higher dimensional HMM models, where Markov states generate sequences rather than a single vector observation. This enables to overcome limitations of the standard HMMs, such as poor duration modeling, assumptions on conditional independence and restrictions imposed by frame-based observations.

3.1. Statistical Trajectory Models

Goldenthal [4] has demonstrated that the STM represent a viable alternative in ASR, capable of achieving virtually identical performance as that of the state-of-the-art HMMs under low-noise conditions. STM is a segment-based approach capable of capturing the dynamical behaviour and statistical dependencies of the acoustic attributes representing a speech waveform.

3.2. Hidden Markov Models

The most flexible and widely known tool for experiments with HMMs is the HTK [9]. We used the HTK version 2.0 in our experimental work. One of the main goals in the comparative performance analysis was to get an insight into the principal ASR paradigms, e.g. what is the difference in performance between *baseline* HMMs and the STM approaches. Therefore, delta and acceleration coefficients were *excluded* from the input speech representation and systems were *not* exhaustively optimized.

A comparison between the STM and HMMs was carried out by applying a HMM architecture optimized in our previous experiments on the OGI_TS telephone bandwidth speech database [3].

4. EXPERIMENTS

HMM and the STM systems were evaluated using three preprocessing methodologies (MFCC, RASTA [5], and FBDYN [1]) running the experiments using two databases,

namely the CTIMIT and the downsampled TIMIT. The parameters for the mel-frequency cepstra (MFCC) were: Hamming window duration 16ms, frame period 5ms, number of output parameters 16 + additional zero'th cepstral coefficient (MFCC_0), analysis order 32. The RASTA coefficients were computed from the mel-frequency cepstra using the filter

$$H(z) = 0.1\, z^4\, \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

FBDYN parameters used identical forward $l_k(n)$ and backward $r_k(n)$ masking Gaussian lifters defined by

$$l_k(n) = r_k(n) = \alpha\beta^{n-1} \exp\left(-\frac{k^2}{2(q_0 - \nu(n-1))^2}\right)$$

where n denotes the time delay and k the order of the cepstral coefficient vector component. The values of $\alpha = 0.08$, $\beta = 0.78$, $q_0 = 18$, $\nu = 1$ were used. Since a 5ms frame period was used, the duration of the forward N_f and the backward masking effect N_b was set to $N_f = N_b = 6$, i.e. 30ms.

The STM system was implemented (for details and definitions of the STM parameters, see [4]) using a trajectory invariant generation function *Traj2*, i.e. fractional linear interpolation with fixed endpoints, proven to be the most efficient in [4]. The tracks had 10 states and the parameter Q (the number of sub-segments in averaging the errors) was set to 3. As the number of input parameters $P = 17$ was used, the resulting joint-Gaussian probability density function statistical model based on the error signal resulted in a dimension of $PQ = 51$.

The HMMs used were 5 state, one skip, two mixture left-to-right models [3]. No grammar was used during the phonemic classification experiments reported in the next section.

4.1. Phonemic Classification Results

The 61 phones used originally in the database were first relabelled into the 38 models used for training, as defined in [4]. The results from the phonemic classification experiments are summarized in Table 1.

The results show that for the CTIMIT the STM achieved at least 10% better phonemic classification than the HMM system. For both systems the RASTA preprocessing yielded the highest recognition scores. FBDYN preprocessing was found to perform poorer than the MFCC and RASTA. A possible explanation for a significantly worse performance of FBDYN using HMMs is that while dynamic cepstra emphasizes the transitions, the chosen number of mixtures was insufficient to model the increased variability in the data.

When comparing these results with the ones obtained on the TIMIT database we see that the absolute differences in performance between the two systems are relatively smaller. It is interesting to note that in this case the RASTA preprocessing yielded worse performance than the MFCC for both the STM and HMM systems. The most efficient preprocessing for the STM system was found to be the MFCC, while for the HMM the dynamic cepstra yielded an improved performance.

CTIMIT			
	STM		
performance	MFCC	RASTA	FBDYN
training set	43.43%	44.43%	38.00%
test set	36.84%	37.87%	32.26%
	HMM		
performance	MFCC	RASTA	FBDYN
training set	24.98%	29.20%	22.80%
test set	23.43%	27.84%	22.29%
TIMIT (8k)			
	STM		
performance	MFCC	RASTA	FBDYN
training set	61.02%	57.39%	53.94%
test set	54.87%	51.08%	48.56%
	HMM		
performance	MFCC	RASTA	FBDYN
training set	53.97%	50.42%	55.10%
test set	52.30%	48.08%	53.45%

Table 1: Performance scores of the STM and HMM systems in phonemic classification on the CTIMIT and downsampled TIMIT databases.

4.2. Confusion Matrix Analysis Results

A rank-order statistical analysis was conducted on the confusion matrices to determine the phonemes which achieved the best classification scores. A general conclusion is that for both systems tested on CTIMIT, the highest performance was achieved on the diphthongs, e.g. /ay, ae, iy, oy/. This is an expected result since in noisy cellular environment phonemes with high spectral dynamics should in principle be easier to recognize and less prone to noise.

On TIMIT, however, phonemes such as /w/ and /f/ also achieved a very good classification performance. This is not surprising since at high signal-to-noise ratio conditions these phonemes exhibit spectral characteristics discriminant enough to be relatively easily recognized.

Detailed results on top five rank-order statistical analysis on confusion matrices are summarized in Table 2.

CTIMIT						
	STM					
	MFCC		RASTA		FBDYN	
rank	phon	% corr	phon	% corr	phon	% corr
1	ay	63.2%	b	76.9%	ae	65.9%
2	ae	59.3%	ay	62.1%	ay	58.4%
3	aa	55.1%	ae	56.4%	s	56.4%
4	iy	52.5%	oy	56.1%	aa	55.0%
5	oy	52.0%	ey	53.8%	ey	54.1%
	HMM					
	MFCC		RASTA		FBDYN	
rank	phon	% corr	phon	% corr	phon	% corr
1	ay	54.2%	oy	58.8%	w	51.3%
2	oy	48.6%	ay	58.4%	ay	47.1%
3	w	45.7%	aw	54.3%	aw	46.5%
4	ae	44.2%	ae	52.2%	ae	46.2%
5	aw	42.6%	ey	50.8%	aa	43.7%
TIMIT (8k)						
	STM					
	MFCC		RASTA		FBDYN	
rank	phon	% corr	phon	% corr	phon	% corr
1	ay	75.2%	b	80.3%	ae	73.0%
2	s	71.1%	w	72.2%	ay	72.8%
3	w	70.1%	y	71.4%	s	69.1%
4	r	68.9%	oy	70.9%	ey	69.0%
5	hh	68.8%	ay	68.0%	aa	67.0%
	HMM					
	MFCC		RASTA		FBDYN	
rank	phon	% corr	phon	% corr	phon	% corr
1	f	71.6%	oy	69.6%	f	72.8%
2	oy	68.2%	ay	64.9%	w	65.4%
3	w	65.4%	w	59.2%	y	65.2%
4	y	63.7%	iy	57.0%	sh	62.2%
5	ay	59.5%	ey	55.0%	ay	60.6%

Table 2: Top 5 performance scores from the confusion matrix analyses of the STM and HMM systems in phonetic classification on the CTIMIT and the downsampled TIMIT databases.

5. CONCLUSIONS

The performance results of the STM approach were found to be significantly better than those of the HMM on CTIMIT and the 8kHz-downsampled TIMIT databases. Neither of the systems was optimized. In phonetic classification experiments on CTIMIT, the STM yielded a 10% improved performance over the HMM when RASTA preprocessing was used. In noisy cellular environment both systems achieved the best performance using the RASTA preprocessing. In general for both systems, confusion matrix analyses revealed that the highest classification results were achieved for the diphthongs.

Given the insights obtained from the phonetic classification experiments, future work will be concentrated on an improved STM modeling for cellular environments and on implementations necessary for recognition experiments.

6. ACKNOWLEDGEMENTS

Bojan Petek gratefully acknowledges a postdoctoral scholarship awarded by the Slovenian Science Foundation. His research was also supported in part by the postdoctoral project Z2-7171-0781-95 granted by the Ministry of Science of Slovenia and a research grant from the Institute of Electronic Systems, Aalborg University, Denmark.

7. REFERENCES

[1] T. Beppu, and K. Aikawa, "Spontaneous Speech Recognition Using Dynamic Cepstra Incorporating Forward and Backward Masking Effect", Proc. Eurospeech'95, pp. 511-514, 1995.

[2] K. L. Brown, and E. B. George, "CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition", Proc. ICASSP'95, pp. 105-108, 1995.

[3] P. Dalsgaard, O. Andersen, H. Hesselager, and B. Petek, "Language Identification Using Language-Dependent Phonemes and Language-Independent Speech Units", Proc. ICSLP'96, pp. 1808-1811, 1996.

[4] W. D. Goldenthal, *Statistical Trajectory Models for Phonetic Recognition*, PhD Thesis, MIT/LCS/TR-642, September 1994.

[5] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP Speech Analysis Techniques", Proc. ICASSP'92, pp. 121-124, 1992.

[6] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", Proc. ICASSP'90, pp. 109-112, 1990.

[7] R. G. Leonard, "A database for Speaker Independent Digit Recognition", Proc. ICASSP'84, pp. 42.11.1-4, 1984.

[8] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition", BU Tech. Rep. ECS-95-002, October 1995.

[9] S. J. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK Book*, 1995.