

# PARALLEL MODEL COMBINATION AND WORD RECOGNITION IN SOCCER AUDIO

*Jack H. Longton and Philip J.B. Jackson*

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK.

[j.longton@surrey.ac.uk]

## ABSTRACT

The audio scene from broadcast soccer can be used for identifying highlights from the game. Audio cues derived from these sources provide valuable information about game events, as can the detection of key words used by the commentators. In this paper we interpret the feasibility of incorporating both commentator word recognition and information about the additive background noise in an HMM structure. A limited set of audio cues, which have been extracted from data collected from the 2006 FIFA World Cup, are used to create an extension to the Aurora-2 database. The new database is then tested with various PMC models and compared to the standard baseline, clean and multi-condition training methods. It is found that incorporating SNR and noise type information into the PMC process is beneficial to recognition performance.

*Index Terms*— Audio indexing, soccer, HMM

## 1. INTRODUCTION

With growing distribution of sports footage over a wide range digital media it is desirable, to have more flexible viewing and browsing capabilities. To facilitate this, an index of important events is required but conventional manual annotation is slow and costly, so automatic alternatives are preferable. Soccer also poses an interesting problem because it has a relatively unstructured content and a complex auditory scene with multiple sources. Video techniques have been typically applied to the problem [1] but are unable to describe the a variety of semantics of the game, notwithstanding their high computational cost compared to audio techniques. Audio therefore has a crucial role to play in automatic highlighting and to perform the task audio cue schemes have typically been used to bridge the gap between different levels of information. It is a general approach used consistently among [2] and [3], to bridge low level (audio features) to high level (game events) and helps parse semantic content. Along with such crowd and other audio cues the commentator can also provide extra information, with the emotional content giving a sense of mood and words linking directly to events. A typical system may look for excited speech [4] or keywords linked to specific play scenarios [5]. However, word recognition in soccer audio is a tricky

task due to the level of noise present but provides key links to the type of actions that have been performed. Also, a fully integrated system of the various sources of information has not been implemented. In this paper we focus on the word recognition task.

In a soccer match or any sporting event it can be assumed that the auditory scene is comprised of audio from microphones in different locations. In general we can assume that these include the stadium microphones and studio microphones. If word recognition is the primary concern, commentator speech can be assumed to be one primary stream and the crowd noise another. To tackle such a problem, three well known methods for dealing with noisy speech when using an HMM structure can be used and depending on the level of information that is known about the noise stream a different model may be appropriate. The methods are multi-conditional training, parallel model combination (PMC) [6] and the factorial HMM (FHMM) [7]. Multi-conditional training is simply training the speech models with noisy conditions. However, PMC and the FHMM seek to use separate models for the target stream and background streams, where the PMC simplifies the noise model to a singular state HMM. A factorial HMM is comprised of a set of mutually exclusive HMMs, and to model the crowd and commentator two streams could be used. A model consisting of two HMMs based on Gaussian mixtures, demonstrates the independent evolution of each Markov chain whose Gaussian mixture state representations are combined to form the observation vectors. Such a system is clearly capable of modeling a parallel asynchronous label sequence and is analogous to the commentator crowd mixing. The FHMM does have a crucial reliance on the quality of model combination scheme used, which is trivial when using linear features. However, linear features are rare in audio recognition systems, with nonlinear features such as MFCCs and PLPs used, which compress the audio spectrum for more robust recognition. The compression causes the exact model combination through numerical integration to be computationally intensive, for this reason several approximations have been tried and PMC schemes [6] have been shown to be successful for the improvement of speech recognition in noisy environments.

In this paper we present an empirical study of word recognition in various noise types typically found in soccer matches

and other sports. To facilitate this we extended the Aurora-2 database to include new training and test sets, to assess the effect of various noise types and training schemes on the small vocabulary recognition task. We then assess whether use of extra information such as noise type and SNR when incorporated in the system can be used to improve results.

## 2. MODELLING METHOD AND AUDIO DATA

### 2.1. Aurora-2

The Aurora-2 database [8] is designed for testing speech recognition on noisy data and in particular additive noise. It is a small vocabulary task with sentences containing strings of digits zero to nine. There are three test sets: Set A babble, subway, car, exhibition; Set B restaurant, street, airport, station; Set C subway M, street M. The training set includes 8440 utterances selected from the TIDigits database, with 55 male and 55 female adults forming a clean set of data. The same utterances are then split into 20 subsets, and each combined with four different noise scenarios and 5 different SNRs, forming the multi-conditional training set. The training set contains five different SNR conditions: clean 20dB, 15dB, 10dB and 5dB. The test sets have two extra SNR conditions which are 0dB and -5dB. In test set A the audio has matched channel and noise conditions, thus the same noise types and channel frequency responses are used as in the training data.

### 2.2. Data collection

Data was collected from TV broadcasts of the 2006 FIFA World Cup, as broadcast on ITV and BBC. All audio was recorded with 16 bits at 48kHz in stereo and was captured at broadcast quality and then downsampled to 8kHz to match the Aurora data. Speech segments were then cut out and the noise mixed to create different noise files. Salient examples of each audio cue were extracted from each of the games and combined in a singular audio file. These were then manually checked to remove any inconsistencies that occurred due to discontinuities. The fant tool [9], used in the mixing of audio to create the Aurora database, was then used to mix the soccer data with the speech data.

### 2.3. Noise types

For the purposes of providing an extension to the Aurora database it was necessary for a set of audio cues to be defined. Audio types were chosen on the basis that they provide interesting information about the game events, the reasons for this were simple; noise types provide likely additional information that may be exploited in a larger system and provide a likely set of the most difficult recognition scenarios. In order to create a new training set and test set, four audio types were required to be defined. The noise types included were

crowd whistling, crowd singing, crowd applause and neutral crowd noise. The crowd whistling is defined as a cumulative whistling generally showing dissatisfaction by the majority of the crowd, which resulted in a frequency band of high energy between 3 and 4kHz. Crowd singing was defined by strong peaks around 0.5, 1.5 and 2.5kHz. Crowd applause had a wider spectrum and was defined by its appearance after cheering or a particularly interesting event. A crowd neutral was also defined.

### 2.4. Distortion

Recognition robustness in a speech recognition system is influenced by its ability to cope with distortion. The presence of background noise causes additive distortion. Additionally convolutive distortion can cause an altered frequency characteristic. In this paper only the effect of additive noise is assessed.

### 2.5. Extended database

Using the same structure that exists within the Aurora-2 data a new test set, set D, was constructed from the four noise types: crowd singing, crowd whistling, crowd applause, crowd neutral. Correspondingly a matched training set, was also created.

### 2.6. Reference model structure

For the baseline model structure the same topology of HMM was used with an equal number of recognition units as in the original Aurora-2 scripts, which were used. In speech recognition a conventional feature set consists of 39 features, which included 12 MFCCs and the log energy with the corresponding delta and acceleration coefficients. For the feature set we used, the log energy was discarded, which was swapped for the zeroth MFCC for the purpose of PMC but is otherwise exactly the same as the standard ETSI frontend [10] used for the Aurora-2 benchmarks. HCopy was also used for the feature extraction. The digit models were trained in the same way with the standard Aurora frontend which included: 18 states per word with 2 dummy states at beginning and end, simple left-to-right models with no state skipping, 3 Gaussians per state, only the diagonal of the covariance matrix used. Additionally pause and silence models were used.

### 2.7. Parallel Model Combination

The PMC method is specifically designed with the purpose of modeling the effects of additive noise. One of the key advantages of the technique is that it can be applied in a number of different noise scenarios where only the noise statistics need be known. The methods include the log-normal approximation, Taylor series expansion and the resampling method. Of

model	s	w	a	n	avg
clean	43.4	54.2	46.9	43.0	46.9
multi-conditioned	16.4	23.1	18.4	18.1	19.0
original unmatched	24.0	27.5	21.6	23.4	24.1
PMC 1 mix clean	22.7	24.3	23.6	23.6	23.5
PMC 2 mix clean	21.3	23.7	22.5	22.4	22.5
PMC 1 mix multi	15.0	27.9	15.5	16.7	18.8
PMC 2 mix multi	14.1	25.4	15.1	15.5	17.5
PMC 8 mix clean	21.0	22.8	21.7	21.7	21.8
PMC noise type	21.8	22.1	22.0	22.8	22.2
PMC SNR	13.3	22.4	13.8	15.5	16.3
PMC noise type	21.8	22.1	22.0	22.8	22.2
PMC SNR	13.3	22.4	13.8	15.5	16.3

**Table 1.** Average recognition results (word error rate %) for the different model training schemes tested on test set D and results for individual noise types crowd singing (s), crowd whistling (w), crowd applause (a) and crowd neutral (n).

these methods only the resampling has been applied to delta and acceleration features.

The resampling PMC [6] was used for the model combination. To recalculate the parameters, three separate equations are used. To recalculate the static observation parameters  $x(\tau) = \log(\exp(s(\tau)) + \exp(n(\tau)))$  where  $s(\tau)$  is the generated speech observation at time  $\tau$  and  $n(\tau)$  is the generated noise observation. The equation can be simply extended for the delta and accelerations. To adapt for different SNRs for either the speech or the noise,  $\log(g)$  is added to either  $s(\tau)$  or  $n(\tau)$ , where  $g$  is the desired gain factor of the particular model. If delta or acceleration parameters are to be used, the equations require the storage of the statistics about the correlations of  $\Delta^2 s(\tau - \omega)$ ,  $\Delta s(\tau - \omega)$ , in addition to the standard statistics normally collected in an HMM. However, the statistics of the model at  $\tau - \omega$  can be approximated by those of the time  $\tau$  [6], thus only the standard HMM parameters need to be used, which was the method adopted.

## 2.8. Implementation and software

For testing and training software and scripts from the Aurora project were used. These were used in conjunction with HTK version 3.2 for feature extraction, training and recognition, with HResults used for analysis of the results. The annotation was done through multiple stages where labels were scored from 1 to 3 with 3 being the most salient, the most salient examples were extracted before being rechecked. Code to perform the PMC was written in the Software library RAVL using HTK format HMM definitions.

## 3. EXPERIMENTAL RESULTS

The recognition results presented in this section consist of the application of various modeling schemes, based on the Au-

roras baselines and the application of the PMC. The results are listed as word error rates (WER) in table 1. The WERs are averaged over the whole range of SNRs present in the test set.

The two baseline models trained for Aurora speech recognition task are models trained on clean speech and models trained on multi-conditioned speech. Using the standard Aurora training data, the performance of the clean speech model on test set A was 46.0% and the multi-condition trained models 21.6%. Using the new training set and test set D Similar performance was witnessed, with the models trained on multi-conditioned data (19.0%) outperforming those trained on clean data (46.9%). A comparison of the multi-conditional trained models was also made when tested on unmatched data. The models trained using the new training set were tested on test set A and those trained on the original training set tested on test set D. The performance of the unmatched models in both scenarios led to a higher word error rate, which was 24.1% instead of 19.0% for test set D. However only marginal performance losses were seen at higher SNRs (10dB and above), thus even fairly unmatched models may still provide inherent robustness.

To compare the effectiveness of the PMC with the baseline results, a one mixture Gaussian noise model, was considered. Applying only a single mixture Gaussian kept the model complexity the same after the application of the PMC. The single mixture was trained over the features for the various SNRs and noise types used in the matched data. Using the PMC method the clean and multi-conditional trained HMM models were adapted using the simple noise model, creating two new model schemes. Improvements were seen with the clean speech model PMC, with a 23.5% WER, however it was worse than the baseline. Overall the multi-conditional speech model PMC performed best, with an overall 18.8% WER, but improved only slightly (0.2%) upon the baseline. The models created by PMC, with only a singular mixture Gaussian noise model gave limited performance. A singular Gaussian may not capture the variety or structure of the intended target noise. To improve the PMC, a 2 mixture Gaussian noise model was used. The noise model was then combined via PMC with the clean and multi-conditioned speech models. For the clean speech PMC the WER improved to 22.5% and for the for the multi-conditioned speech the accuracy improved to 17.5%. Whilst the use of multi-conditional speech models had a positive effect for three of the noise types it has the opposite for the crowd whistling with a WER of 25.4% compared to the baseline of 23.1%. Also whilst there was an improvement at the lower SNRs, performance at the higher SNRs decreases particularly the performance of clean speech up from 1.6% to 5.8%. In contrast the change at 20dB SNR was from 3.3% to 3.7%. Further, an 8 mixture Gaussian was trained for the noise model and combined with the clean speech model. The 8 mixture model did outperform the 2 mixture model with the reduction in WER from 22.5% to 21.8%.

Four different model schemes have been applied in different scenarios. It has been shown that using multi-conditioned training data improves the accuracy of word recognition, which is still robust when using unmatched data. Also, two PMC schemes have been assessed, with the PMC trained on the multi-conditioned speech models performing best however performance is noise specific and clean speech recognition is degraded. The multi-conditioned speech models thus give the most rounded performance but are less suited to low SNR scenarios.

In a large scale system information about the audio streams may be known or estimated, such as noise type or SNR. Thus an oracle that could predict the noise type and SNR was assumed, to assess the impact of such information. Thus, SNR and noise type specific experiments were designed. For the SNR experiment, a 2 mixture SNR specific noise model was trained for each SNR and combined with the clean speech models. For the noise type experiment a 2 mixture Gaussian was trained for each of the four noise types. The resultant noise models were combined via PMC with the clean speech models. The overall accuracy of the SNR specific experiment was 16.3%, beating the baseline. The improvement however was dominated by those at low SNRs, whereas the baselines still performed slightly better in the mid ranges. The noise type experiment yielded less impressive results with the overall accuracy of 22.2%. The noise type experiment did perform relatively well for the crowd whistling condition achieving the best WER of 22.1% showing the benefit prior knowledge of the noise type in this case. In addition the separate HMMs for the SNR and noise type specific experiments were run in parallel, and the log likelihood used to select the most probable result. Removing the oracle had no effect with the same performance achieved.

The effects of recognition with different noise types varies. With the four types used in our test sets clearly the performance is by far worse when considering crowd whistling. The explanation for this is that the spectra of the other 3 types are more evenly spread across the frequency range. The whistling however is concentrated mainly at higher frequencies. This has little impact on the recognition of most of the digits but drastically effects the recognition of the number six, which also has strong similarities, giving numerous insertion errors.

#### 4. CONCLUSIONS

In this paper, the application of various model training and recognition schemes has been described, with a particular focus on parallel model combination. From the range of models tested it is apparent that for the noise types defined that multi-condition trained data is very effective at producing good recognition results for medium to high SNRs. However as the SNR decreases to low levels such a method struggles. By using a PMC scheme similar performance at the mid to high

SNRs was achieved coupled with improvement at low SNRs. For most of the audio types, noise type information was not that important in designing better models for recognition, however for the crowd whistling there was a clear benefit of including this information in the system. However estimation of the SNR was a strong factor in improving recognition accuracy.

#### 5. REFERENCES

- [1] Chung-Lin Huang, Huang-Chia Shih, and Chung-Yuan Chao, "Semantic analysis of soccer video using dynamic bayesian network," *IEEE - Trans. Multimedia*, vol. 8, pp. 749–760.
- [2] Min Xu, N. C. Maddage, Changsheng Xu, M. Kankanhalli, and Qi Tian, "Creating audio keywords for event detection in soccer video," in *Proc. ICME*, Washington, DC, USA, 2003, pp. 281–284, IEEE Computer Society.
- [3] Mark Baillie and Joemon M. Jose, "An audio based sports video segmentation and event detection algorithm," in *Proc. CVPRW*, Washington, DC, USA, 2004, p. 110.
- [4] Hyoungh-Gook Kim, Steffen Roeber, Amjad Samour, and Thomas Sikora, "Detection of goal events in soccer videos," in *IS and T/SPIE's Electronic Imaging 2005*, San Jose, CA USA, jan 2005.
- [5] M. Sano, I. Yamada, H. Sumiyoshi, and N. Yagi, "Automatic real-time selection and annotation of highlight scenes in televised soccer," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 224–232.
- [6] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," in *IEEE Transactions on Speech and Audio Processing*, 4(5):352–359, September 1996.
- [7] Zoubin Ghahramani and Michael I. Jordan, "Factorial hidden Markov models," in *Proc. NIPS*, David S. Touretzky, Michael C. Mozer, and Michael E. Hesselmo, Eds. 1995, vol. 8, pp. 472–478, MIT Press.
- [8] Pearce D Hirsch, H.-G., "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ASR*, Paris, France, 2000, pp. 181–188.
- [9] "<http://dnt.kr.hs-niederrhein.de/download.html>," .
- [10] ETSI standard document, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm," in *ETSI ES 201 108 v1.1.2 (2000-04)*.