

NLP Project Report

Group 13

Saranya .M .S (CS15D006), Sridharan Sankaran (CS14S024)

November 27, 2015

Project title

Weather/Cyclone forecast using NLG techniques

Problem Definition

In this project work, a weather report is generated to forecast the climatic conditions and cyclone warnings from the satellite images using image processing, data mining and Natural Language Generation techniques in a conversational language instead of telegraphic-weather-language. A text-to-speech synthesis system is used to read the report using synthesized voice. The navigation direction of cyclone is predicted after processing the satellite images.

The forecast is not only made for the cyclone warnings but also for forecasting daily weather conditions. The location of the cyclone is given with respect to the near-by cities for easy readability. A look-up table with list of near-by cities along with their latitude and longitudinal values is used for this purpose. The trajectory of the cyclone is also calculated based on the direction and speed of the cyclones along with the nearby target cities. A 5-day weather forecast is made with the time-series of various weather parameters over a period of 5 days given by an open weather API. Then a text to speech synthesis system is implemented using HMM based speech synthesis system to read this text report for visually challenged people.

1 Introduction

Weather forecasting become an essential thing in our daily life to know the upcoming natural calamities. With the advancement in the meteorology and communication technology, images obtained by the satellites stationed above are used in weather forecasting. But the forecasting requires intense training and advanced education to interpret the information form the satellite pictures. Meteorological department has already automated the gathering of satellite images in the database. Further advancement can be brought by automating the weather forecasting. Although the automated forecast may not be as accurate as the experts interpretation [1], Forecast Production Assistant (FPA) was developed to automate the routine aspects for public health and safety in hazardous weather [3].

2 Input sources

Two types of input data are required to forecast the weather report. For cyclone forecast, satellite images are needed and for daily weather forecast other weather parameters like rain, wind, humidity etc are needed. The source of these data-sets are given in the subsections below.

2.1 Image database

The satellite images are gathered from Hong kong Observatory website located at [http : //www.hko.gov.hk/wxinfo/intersat/satellite/sate.htm](http://www.hko.gov.hk/wxinfo/intersat/satellite/sate.htm). Here we consider deep convection weather images of Asia taken from satellites at intervals of about once every half-hour. Deep convection clouds are very dense with water content and are seen as red patches in the images. The sample satellite image used for cyclone forecast is shown in the Figure 1.

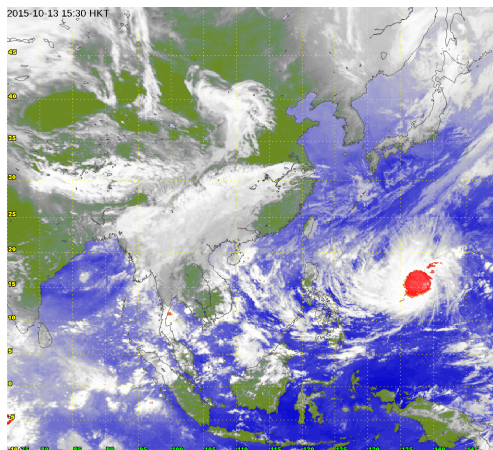


Figure 1: Satellite image of a cyclone

2.2 Weather parameters

For preparing the 5-day weather forecast report, the weather parameters are collected from *www.openweathermap.org* using an API. This site will have the information of atmospheric changes for every 3 hours as shown in the Figure 2.

```
[from] => Mon, 23 Nov 2015 00:00:00 -0600
[to] => Mon, 23 Nov 2015 03:00:00 -0600
[symbol] => moderate rain
[windDirection] => East-northeast
[windSpeed] => Gentle Breeze
[windMps] => 3.92
[temperature] => 26.76
[humidity] => 100
[clouds] => broken clouds
```

Figure 2: Sample weather parameters

2.3 Speech corpus

For developing a TTS system, we need the recorded speech content, generally called as speech corpus which serves as a database. The text-to-Speech synthesis process needs two types of data namely, 1) Text data and 2) Speech data. The former one is the text form of the content which is to be recorded to get the later form of data. Here for the purpose of developing a phoneme-based speech synthesis model of English language, one hour of CMU-artic corpus is used. The speech utterances in this database is recorded in studio environment with the sampling rate of 16KHz and 16 bit quantization.

3 Proposed weather report generator

Natural language generation (NLG) techniques are used to generate the weather forecast report. This process includes three phases as follows

- Preprocessing
- Weather report generation using NLG
- Speech synthesis using TTS

3.1 Preprocessing

The preprocessing of raw data is always a mandatory step to prune the unwanted information.

3.1.1 Image preprocessing

The satellite image obtained here has a wide range of geographical information along with cyclone. The preprocessing step filters out those unwanted image features and retains the cyclone alone.

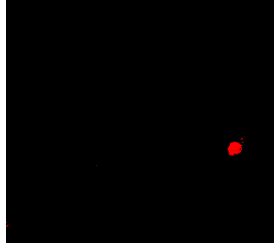


Figure 3: Preprocessed cyclone image

The cyclone in the image is considered as a cluster of pixels and the centroid is treated as *"the eye of the cyclone"*. The centroid is identified using *"K-means clustering algorithm"*. A database of prominent cities along with its position in the image is maintained separately. This is used as a look-up table (Section 3.1.1) to decide the current location of the cyclone and its trajectory. The rate of change of cyclone path is also used to calculate the speed and direction of the cyclone. The whole algorithm is explained in Algorithm 1

Algorithm 1 Cyclone tracking algorithm

- 1: **Input:** Images from HK observatory at 30 minutes interval
Output: Weather report as English language text
 - 2: Process input image to retain only those pixels corresponding the convection cloud movement
 - 3: Find all the pixel values of resulting image
 - 4: Retain only the red pixels corresponding to the convection clouds
 - 5: Use DBSCAN algorithm to identify only those clusters (cyclones) that are big enough to be reported
 - 6: Use k-means to find the centroids of these clusters (cyclones)
 - 7: Track the movement of these centroids from one image to another as a time series data
 - 8: Create weather report that locates the movement of the cyclone along with its speed using nearby city database lookup table
-

Look-up table

Generally weather prediction is globally classified into different sectors like East pacific, West pacific, Indian basin, Atlantic basin etc. Each sector will be monitored by a set of geo-stationary or mobile satellites. So each satellite may cover

only limited area. The major cities falling under those sectors can be easily enlisted based on their latitude and longitude values.

Since our database is around South-East Asia , the major cities and places in and around are identified and tabulated offline. This table acts a look-up table to include the city names and warn the people of those locality. The near-by victim of a hazardous climate can be calculated using the *Euclidean distance* between the latitude and longitude of the cyclone and that of the cities listed in the look-up table.

3.2 Text preprocessing

Using weather API, the forecast of various weather parameters like wind direction, speed, temperature, humidity etc are collected in an interval of 3-hours as a time-series data over a period of five days. The data is of the format shown in the Figure 2. This data is preprocessed to get a parameter table for easy accessibility as shown in the Figure 4.

From	to	symbol	windDirection	windSpeed	windMps	temperature	humidity	clouds
Tue, 24 Nov 2015 15:00:00	Tue, 24 Nov 2015 18:00:00	sky Is Clear	East-northeast	Gentle Breeze	4.12	25.67	100	clear Sky
Tue, 24 Nov 2015 18:00:00	Tue, 24 Nov 2015 21:00:00	few Clouds	East-northeast	Gentle Breeze	3.86	25.36	100	few Clouds
Tue, 24 Nov 2015 21:00:00	Wed, 25 Nov 2015 00:00:00	broken Clouds	NorthEast	Light Breeze	2.87	26.03	100	broken Clouds
Wed, 25 Nov 2015 00:00:00	Wed, 25 Nov 2015 03:00:00	scattered Clouds	North-northeast	Light Breeze	2.91	26.78	98	scattered Clouds

Figure 4: Preprocessed weather parameters table

4 Weather report generation using NLG

The following NLG techniques are used to prepare the weather forecast report.

1. Content determination
2. Discourse planning
3. Sentence aggregation
4. Lexicalization
5. Linguistic realization

$$\left[\begin{array}{l} \textit{Message id : cyc01} \\ \textit{Subject : Location} \\ \textit{Arguments :} \end{array} \left[\begin{array}{l} \textit{City : Taipai} \\ \textit{Distance : 80Kms} \\ \textit{Direction : Nort - East} \end{array} \right] \right]$$

$$\left[\begin{array}{l} \textit{Message id : cyc02} \\ \textit{Subject : Trajectory} \\ \textit{Arguments :} \end{array} \left[\begin{array}{l} \textit{TargetCity : Manila} \\ \textit{Speed : 25kmph} \\ \textit{Direction : Nort - East} \end{array} \right] \right]$$

Figure 5: Message structure formulated by content determination

4.1 Content determination

Content determination is the process of deciding *what information should be communicated in the text*. This includes the identification of unchanging texts, directly available data, computable data and unavailable data. In our scenario, the initiative phrases like *"Today's weather report"*, *"The cyclone"* are unchanging texts. Information like *name of the cities, directions*, and parameters like *humidity, speed of the wind, temperature* which are directly available from the data are classified as directly available data. The *speed of the cyclone* is computed from the movement of the cyclone and hence it is considered as the *computable data*.

The contents are usually described using a set of messages. These messages are data objects used in subsequent processes. The messages will take the structure as shown in the the Figure5.

4.2 Discourse planning

Discourse planning is the process of deciding the order and structure of the message to be conveyed. Good structuring increases the readability of the text thereby increasing the sense to be conveyed considerably. Hence the order of information to be presented must be organized in a simpler and natural way. In our scenario, the cyclone will be reported first followed by the usual weather forecast. While reporting the cyclone, its current location will be reported first with reference to the nearby city followed by the trajectory and the cities in the trajectory path. We mention the closest city in the trajectory as the "target city". Likewise, in the 5-day weather forecast report, the initial phrase is followed by the outlook of the day (normal/rainy/hot) along with the nature of the wind and temperature.

4.3 Sentence aggregation

Sentence aggregation is the process of grouping messages together into sentences (for example, many short sentences or a few longer sentences). In forecasting the cyclone information, the messages shown in the Figure 5 are represented in two individual sentences as shown below.

Example 1: *There is a cyclone currently located at 80 Kms North-East of Taipei. It is likely to move towards Manila in North-west direction at the rate of 25 kmph.*

Similar kind of aggregation is done among the 5-day weather messages and two sentences are formulated as given below.

Example 2: *Today will be a normal day with few clouds and gentle breeze. The temperature will be around 27°C.*

Example 3: *Today will be a rainy day with overcast clouds and strong wind blowing in South-West direction with a speed of 15 mph. The temperature will be around 27°C.*

The normal weather report based on the 5-day forecast is classified generally into three categories namely *normal*, *rainy* and *hot* as mentioned in the Section 4.2. The classification of days is done mainly based on the attributes namely clouds, temperature and symbol. The basic class information is given in the Table 1

Out-look of the day	Attribute	Status
Normal	Cloud	few clouds clear sky scattered clouds broken clouds
	Temperature	around 25 to 30°C
	Symbol	Sky is clear few clouds
Rainy	Cloud	-
	Temperature	-
	Symbol	light rain moderate rain heavy rain
Hot	Cloud	few clouds clear sky scattered clouds broken clouds
	Temperature	more than 35°C
	Symbol	-

Table 1: Classification of a day's out-look based on weather attributes

4.4 Lexicalization

Lexicalization is the process of deciding which specific words and phrases should be chosen to express the domain concepts. In the case of cyclones the certain phrases like "currently located" is used to give the current location of the cyclone. The phrase "likely to move towards" is used to mention the nearby or next target city that is expected to be affected by the cyclone. In the case of daily weather report, apart from the out-look of the day, phrases like "with a speed of", "today will be a ", "will be around" are used as connectors to connect the attribute values and produce a readable sentence in conversational English.

4.5 Linguistic realization

Linguistic realization is the process of applying the rules of grammar to produce a text which is syntactically, morphologically and orthographically (spelling, hyphenation, capitalization, word break, emphasis, and punctuation) correct. In our scenario we decided to use conversational English with proper English grammar. Thus separate grammatical rules are not formulated for this purpose. We used normal English language grammar to synthesis the sentences by including some grammatically correct connector phrases with proper orthographic notations as mentioned before.

5 Text-to-speech synthesis

The process of generating text report from the image sequence itself in a NLG process. The text report is then read aloud using Text-to-speech (TTS) module, which is another NLG process in order to help the visually impaired people. Given a text in one language, a Text-to-Speech (TTS) system is expected to produce high quality speech signal (corresponding to the text provided) which is intelligible to a human listener. TTS is a well developed area which can be achieved in more than one way. Numerous research work had been carried out in the past to build a high quality speech synthesizer for one language. Speech synthesis is the process of introducing engineering approaches for "talking" machines that uses a sequence of sub-word units. The use of such elements provide flexibility for vocabularies as required for applications like text to speech. This speech synthesis can be carried out in two different ways namely, (1) Unit Selection based speech Synthesis (USS) [2] and (2) Statistical parametric speech synthesis technique [7].

Text-to-speech processing involves three phases as mentioned below

- **Text preprocessing:** This step translates the text string of characters into a new string with resolved ambiguities.
- **Text to phonetic translation:** The processed text is parsed to determine phonetic structure. A sequence of modifications to canonical pronunciations is typically encoded in a series of rewrite rules.

- **Speech synthesis:** Given the sequence of quantities like pitch, duration, amplitude along with unit label, the signal processing component generates speech.

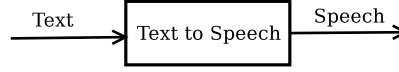


Figure 6: Text to speech synthesis process

5.0.1 HMM-based speech synthesis

Hidden Markov Models (HMMs) can be used to model any time series. HMM tool kit (HTK) developed by Microsoft and Cambridge University is used to build HMMs. Initially HTK was designed for building HMM-based speech processing tools, in particular recognizers. As shown in the picture above, there are two major processing stages involved. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools.

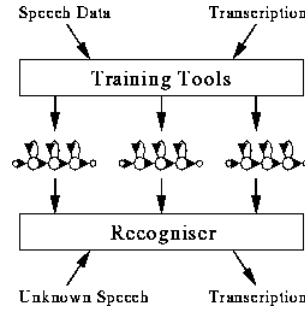


Figure 7: Overview of HTK

The complete process of HMM-based speech synthesis system, shortly called as HTS (H-Triple-S) for a phoneme-level system is explained in the Algorithm 2.

6 Evaluation Measures

The first process of this project are image processing which analysis the image sequence of climatic conditions and creates a weather report as text document with the location, direction, and speed of the cyclones. The accuracy of

- the cyclone location
- the direction of the cyclone movement

Algorithm 2 Steps for text-to-speech using HTS

- 1: **Input:** Sentence(s) as text
 Output: Synthesised speech
 - 2: Prepare the speech corpus.
 - 3: Pre-process the text data.
 - 4: Segmentation and labelling of audio segments in corpus.
 - 5: Manual correction of segmentation.
 - 6: Preprocessing to get the phoneme labels.
 - 7: Model training in HMM Training phase.
 - 8: Verification of trained models.
 - 9: HMM synthesis phase.
-

- the speed of the cyclone movement
- identifying the near-by or next hit areas in terms of city names

are to be evaluated. This can be done by comparing the manually prepared report and the report generated by the automated process for the validation image set.

The second part of the project is to generate the synthesised voice from the text report. The synthesised voice has to be evaluated for

- intelligibility and
- naturalness.

This can be evaluated by a measure called "Mean Opinion Score" (MOS) [6]. This score is calculated by playing the natural human speech and the synthesised speech alternatively to a set of naive speakers and asking them to give scores in the scale of 1 to 5 to measure the intelligibility and naturalness of the synthesised speech. The scale of 1 to 5 has the standard meaning as given below.

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 2: Mean opinion score (MOS)

6.1 Result of TTS

The synthesized sentences are played along with the human spoken to 10 different naive listeners and they are asked to evaluate the sentences based on

MOS in the scale of 5 as shown in the Table 2. The average MOS of 3.4 out of 5 explains that the synthesized voices are intelligible with certain amount of naturalness in it.

6.2 Result of Image processing and NLG

The trajectory of the cyclone is formed by the movement of its eye. The correctness of the target cities and the cities reported as to be encountered in the cyclone path are verified manually.

As per the reference to the NLG papers from the literature, the evaluation of generated weather reports is based on the questionnaire (Figure 8) evaluated by different subjects.

	(a)	(b)
Which text is easiest to read?	<input type="checkbox"/>	<input type="checkbox"/>
Which text is most accurate?	<input type="checkbox"/>	<input type="checkbox"/>
Which text is most appropriate?	<input type="checkbox"/>	<input type="checkbox"/>

Figure 8: Sample questionnaire

We chose the conversational English language to generate the report. So for the comparison and evaluation of the proposed system we used the manually written text reports in the telegraphic-weatherese format as mentioned in [5] and [4]. The telegraphic-human generated sentences and the system generated sentences are shuffled and given to the users and are asked to evaluate the best understandable sentence. The proposed system in conversational format is evaluated as the better system.

References

- [1] E. Goldberg, N. Driedger, and R. Kittredge, “Using natural-language processing to produce weather forecasts,” *IEEE Expert*, vol. 9, no. 2, pp. 45–53, April 1994.
- [2] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, May 1996, pp. 373–376 vol. 1.
- [3] L. Melcuk, “Meaning-text models: A recent trend in soviet linguistics,” *Annual Revision of Anthropology*, vol. 10, no. 5, pp. 27–62, 1981.
- [4] E. Reiter and R. Dale, *Building Natural Language Generation Systems*. New York, NY, USA: Cambridge University Press, 2000.
- [5] E. Reiter, S. Sripada, J. Hunter, and I. Davy, “Choosing words in computer-generated weather forecasts,” *Artificial Intelligence*, vol. 167, pp. 137–169, 2005.
- [6] M. O. S, “<https://en.wikipedia.org/wiki/mean-opinion-score>.”
- [7] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.