Challenges of linking statistical data and phonetic pronunciation software.
Case study: problem of regular statistics establishments' frames in Egypt.

Nehall Ahmed Farouk Mohamed (nehall ahmed@capmas.gov.eg)
Research, computer, and sampling specialist in CAPMAS, Egypt.

Abstract

**Key words:** Recent Neural Networks (RNN), International Phonetic Alphabet (IPA), Master aggregated frame, Automatic Speech Recognition (ASR).

1. **INTRODUCTION**

Different types of statistical data are processed for various reasons to improve the statistical work and to provide new indicators. Some types of these data are measurable, comparable, and linkable but others are not. Statistical work might have a lot of challenges of mixing, comparing, and linking data, these challenges results from the nature of data type.

Many countries face the problem of linking data with each other because it's difficult to be compared. Methods, software, and techniques are implemented to solve this problem. The paper discusses the problem of linking certain case of data in Egypt and developing a methodology or technique on the basis of phonetic system for it. Also the paper discusses the nature of the Arabic language writing in Text To Speech software system (TTs). The case study problem appears in the implementation of the aggregation process for establishments' different frames in CAPMAS to create a master aggregated frame.
2. METHODS

2.1 First: finding out different Problems of linking statistical data that depends on the data type.

2.2 Second: understanding the basics of Phonetic pronunciation software systems and the latest achievements in Arabic language in both (TTS-ASR).

2.3 Third: The nature of Arabic language writing and its challenges for TTS software:

a. Writing and pronunciation of Arabic are Very difficult.

b. Arabic has a lot of problems to be implemented in TTS software.

Fourth: case study :( problem of regular statistics establishments frames' in Egypt).

Central Agency of Public Mobilization and Statistics (CAPMAS, EGYPT) conduct many different regular establishments surveys, each survey has its own frame. CAPMAS seeks to generate a main aggregated frame for all of the regular statistics establishments' frames. The total number of the related overlapped frames is 67 frames. The problem appears in the implementation of the aggregation process because there is no way to compare and link the same establishments. The paper discusses a new developing technique to link the data of the frames (establishments) and generate the aggregated frame.

The current situation of the frames is illustrated as following:

- Different frames of the establishments are overlapped and same establishment exists in different frames.
- All establishments have no unique ID number to be used in data linking.
- Disability of matching the same establishment in the related frames as it is not completely compatible in name but partially compatible because of the nature of writing in Arabic.
- Disability of matching the same establishment in different frames as it exists with different names (about 20% of the frames).
So frames aggregation and unification process is not accomplished due to lack of matching techniques. The idea of linking data here will depend on phonetic pronunciation software technique as a main part in the aggregation process to compare the data first and then linking it.

- The aggregation process will include main steps that are:
  1. Determining and collecting metadata about all of the overlapped related frames.
  2. Determining relationships and inter-relationships between the frames.
  3. Classifying the frames:
     - Relationship (master frames - related frames - independent frames)
  4. In parallel: (Creating a unique ID number- compare through the pronunciation phonetic system).
  5. Final aggregation process (matching through TTS software).
3. RESULTS

The expected results of generating the master aggregated frame will have many effects in our statistical work, economic and technical systems like:

1. Data about one establishment will be collected once.
2. Reduce the fieldwork cost.
3. Helping in generating the administrative data for establishment.
4. Excluding some surveys and affects the total cost.

4. CONCLUSIONS

Linking incomparable data can be achieved by the analysis of the data. The step of finding out relations between different files of data and how to compare then is the most important point to link data. So statisticians must study the nature of data and then think of how to use the most technological systems or methods to link it. Also that phonetic software are useful in comparing and linking data if the suitable software was developed.

REFERENCES

1] Judith Rosenhouse, Arabic phonetics in the beginning of the third millennium.
3] yasser Hifny, Shady Quranny, Salah Hamid, Mohsen Rashwan, and other participation, ARABTALK – An implementation for Arabic Text To Speech system.