CHAPTER 9

CONCLUSIONS
Chapter 9

CONCLUSIONS

The thesis presents work done by the author in the field of Speaker Identification. As detailed in chapter 2 there have been a large number of research workers interested in this problem. The author is benefitted greatly by trying to understand the reported literature. The methods are presented for speaker identification. The classical model for speech production has been used in many approaches to the problem. However, it can be noted (as shown in chapter 2) that the concept of using certain 'characteristics' (later called as features) existed even before the model was presented and universally accepted. Some of these features exist even now in the later works. Altogether, the differences between two speakers have been presented as differences between two networks. The intuitive feeling one gets at this stage is that the learning process, acquiring accents, securing imperfections etc., have not been reflected in the model at all. For instance, it was presented (in chapter 6) that children and pets learn to discriminate between voices. The capacity to discriminate grows with age and expanding areas of activity. It was the intention of the author to include this aspect in this study. This was possible only by proposing a model different from classical model. Also, the model has to depend on a knowledge base to
enable gradual learning. The model, at the same time, has to retain the 'network' properties of the classical model not merely because the model has been extensively and successfully used, but also because certain features are dependent on 'filter' parameters and do not entirely need the knowledge base. It was with this aim that the FTPM model was postulated. Defining and measuring the parameters of the knowledge base turned out to be a tough problem. It was felt that certain irrelevance will creep into the work if greater attention is devoted to the knowledge base and its parameters. Indeed the work would possibly look like one on knowledge engineering and will definitely fall beyond the capability of the author. The model presented finally in chapter 6 is chosen after trying a number of such models proposed and discarded either because all the aspects of learning-knowledge base interaction could not be presented effectively or due to extreme complexity. However, the final model presented is not perfect since all the aspects presented in the heuristics have not found a place in it. Simplicity of implementation was another constraint which led to the choice of the model and methods. As a matter of fact, it was strongly felt that the implementation must be PC oriented. Therefore, a suitable, simple and reasonably good measures have to be devised. The measurements were hence localized on entropy and information content. The measurements proved to be multi-dimensional and the reduction was essential. Though certain
similarity exist with Markov Model and particularly Hidden Markov Model, the difference lies in the definition of probabilities of occurrence of the phonemes. The FTPM model however, did not provide results from the second and near conventional method. The reason for this is due to the fact that the entropy measure defined and used is probably as good or as bad as a speaker dependent measure like Pitch, ACF, Energy etc. As was pointed out in the chapter 8, the performance did not deteriorate too. The model therefore, needs certain changes to include measurable, speaker dependent entropies. The changes are mainly in defining the rate of change (corresponding to transition probabilities in HMM) more precisely but without losing the speaker dependence. This is very difficult to achieve. One has to precisely define and mark a unit within the speech samples. The author has chosen a set of phrases which try to reflect as far as possible the naturalness of the speaker without constraining the actual mode of delivery. Thus it becomes a trial and error method and has taken many trials and time. As a matter of fact, even the choice of phrases took considerable time and as pointed out in chapter 4, some changes had to be made from a set of selected phrases, after study. These two areas, it is felt, can be further explored. Though the FTPM and the new knowledge base model did not show remarkable results, there is potential in it to do better than the other models. A fine tuning of the learning parameters and the way of measurements will improve the
performance. Similarly the phrase set chosen is in English. The subjects are from a background in different language but all belonged to one region. It is likely that if the phrases were in the language where the influence of learning is much more pronounced, for example Malayalam, the performance would have been much better. The reason for this is the fact that English is acquired by all the subjects as a spoken language in school and college thus providing similarity in learning for all the subjects. The error percentages and cross-identification are likely to improve. However, a phrase set independent of drawbacks is almost always impossible to choose. Choice of a phrase set, as of is not based on rigid mathematics. Further work will be interesting.

An attempt was also made to use the concept of eigen frequencies. However, this was not found universal and hence could not be incorporated into the model. The possibility of including this is an open area for further work.

Regarding the system fabricated as front end, it was attempted to duplicate as nearly as possible, toll quality speech. The aim can be furthered to consider systems accessible via telephone.

With regard to the set of new features defined, for example, Symmetry Check method for the determination of pitch,
it was found very effective and consistent. In fact all measurements on voiced portions were found to be consistent not only by this author but also by many others. Why an unvoiced sound should not show consistency in some measure is not known. The author has tried to utilize some unvoiced sound portions also in this work and the performance is not unduly deteriorated. It would be interesting to define some measures for unvoiced sounds, at least a selected few, and explain the lack of consistency in the measures or otherwise.

Yet another interesting aspect is cross-identification. Even in the simple system presented in this work there is a one way consistency of cross-identification. This suggests a possible similarity in the learning environments of the two speakers as measured by the system. However, the lack of reciprocity suggests that the measures are measuring only one aspect of the feature. It is likely that this vector (space) is anisotropic. Many analyses were attempted but were not acceptable. This aspect of non-reciprocity in identification is also an open problem for further work.

The thesis attempts therefore, to present a low cost PC based speaker identification system and in implementing this a few new concepts have been introduced. The performance is not exemplary but neither is it too poor to be discarded.
For a simple system, indeed the performance is comparable to many more sophisticated systems. It may appear that too many features have been selected, but, by a judicious knock-out method and by monitoring the performance it was possible to maintain a good score. However, the knock-out procedure is one of trial and error. There has not been a rigorous analysis of the technique in literature and from the author's experience it is felt that, it is not possible to provide a rigorous analysis or a procedure for this. It is for this reason that one cannot claim that the features selected by the author are the optimum features. It would an interesting further work if optimality in feature selection is defined and proved for a set of features. In the present work the single set of measurements on the samples yields all the features defined and used in this work. To this extent, that is minimizing the data base and computations, this work is near optimal, eventhough the system is not in real time.