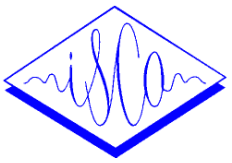

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A UNIFIED APPROACH FOR SPEECH SYNTHESIS AND SPEECH RECOGNITION USING STOCHASTIC MARKOV GRAPHS

Matthias Eichner, Matthias Wolff, and Rüdiger Hoffmann
Dresden University of Technology
Laboratory of Acoustics and Speech Communication
D-01062 Dresden, Germany

ABSTRACT

With the progress of speech synthesis towards the development of complete TTS systems, the databases of speech synthesizers obtain more and more similarity with databases of speech recognizers. This offers new possibilities in combining systems for speech synthesis and recognition. In a new project, we are developing a speech dialogue system with the synthesis and recognition components using unified databases. In this paper, we describe the aim of the project, the framework of the system and first results at the acoustic and the word level.

1. MOTIVATION

With the progress of speech synthesis towards the development of complete TTS systems, the databases of speech synthesizers obtain more and more similarity with databases of speech recognizers. The simple database of early synthesizers consisting of a collection of speech segments described in a suited manner, was complemented by pronunciation dictionaries and a system of rules for generating the prosody. This process of convergence which is not finished at all seems to offer new possibilities in combining systems for speech synthesis and speech recognition. Considering this tendency in a new project, we are developing a speech dialogue system with the synthesis and recognition components using unified databases. The project continues our activities in developing the multilingual TTS system DreSS [1].

The project is aimed to improve our capabilities in education as well as in research:

- **Research**: We are going to develop an experimental tool to improve the insight in the interrelationship of synthesis and recognition during the dialogue act. This very general goal includes a number of interesting details which will be discussed below (cf. 3.).
- **Education**: It is not only necessary to understand but also to explain the speech dialogue process better. Web based methods are the ideal support for teaching speech technology. With our project, we want to extend our activities which started with a web-based project in teaching speech synthesis [2] to a larger area. This includes some additional problems in software technology which are discussed in [3].

This paper describes the current status of the project. Until now, we developed the "acoustic component" which is limited by the phoneme level.

2. PROPOSED FRAMEWORK

The framework of our system is shown in Figure 1. On the left-hand side, from bottom to top, the speech recognition process is shown. The speech synthesis path is represented by the right-hand side, from top to bottom. At each level of processing, the components use common databases.

The modular software structure of the framework enables flexible application of different algorithms. The algorithms which were selected for the first version will be discussed below.

During the recognition process, prosodic information is separated from the data flow. It is known that this information may be very helpful in improving the recognition result. In the case of synthesis, this information has to be added again. This process is crucial for the naturalness of the synthesized speech. Because of these reasons, the modules that are dealing with intonation and time structure will act as essential control elements of the framework.

3. RESEARCH GOALS

3.1 Speech Recognition

Speech recognition is frequently stated to be a solved problem. Many systems show impressive results indeed. Compared with the human recognition performance, however, recognizers show worse results especially in non-laboratory environments [4]. This situation has not changed during the last few years. Recently, we

conducted a detailed performance analysis of 10 state-of-the-art German dictation systems [5] for a computer journal. The investigation resulted in an overall word recognition rate of 85.6 % average with only low differences between the systems.

The question why speech recognizers produce errors is not answered satisfactory until now. This is due to the complexity of the systems which makes tracing of observed errors a complicated task. One of the few available analyses is given in [6]. Even in this investigation, a considerable part of errors (21.5 %) cannot be interpreted.

The method of "Analysis by Synthesis" may be used to evaluate the performance of the recognition process in detail. This is the main goal of our system. It allows an inverse function of the recognizer which will contribute to an optimum selection of feature extraction methods, word and language models, etc.

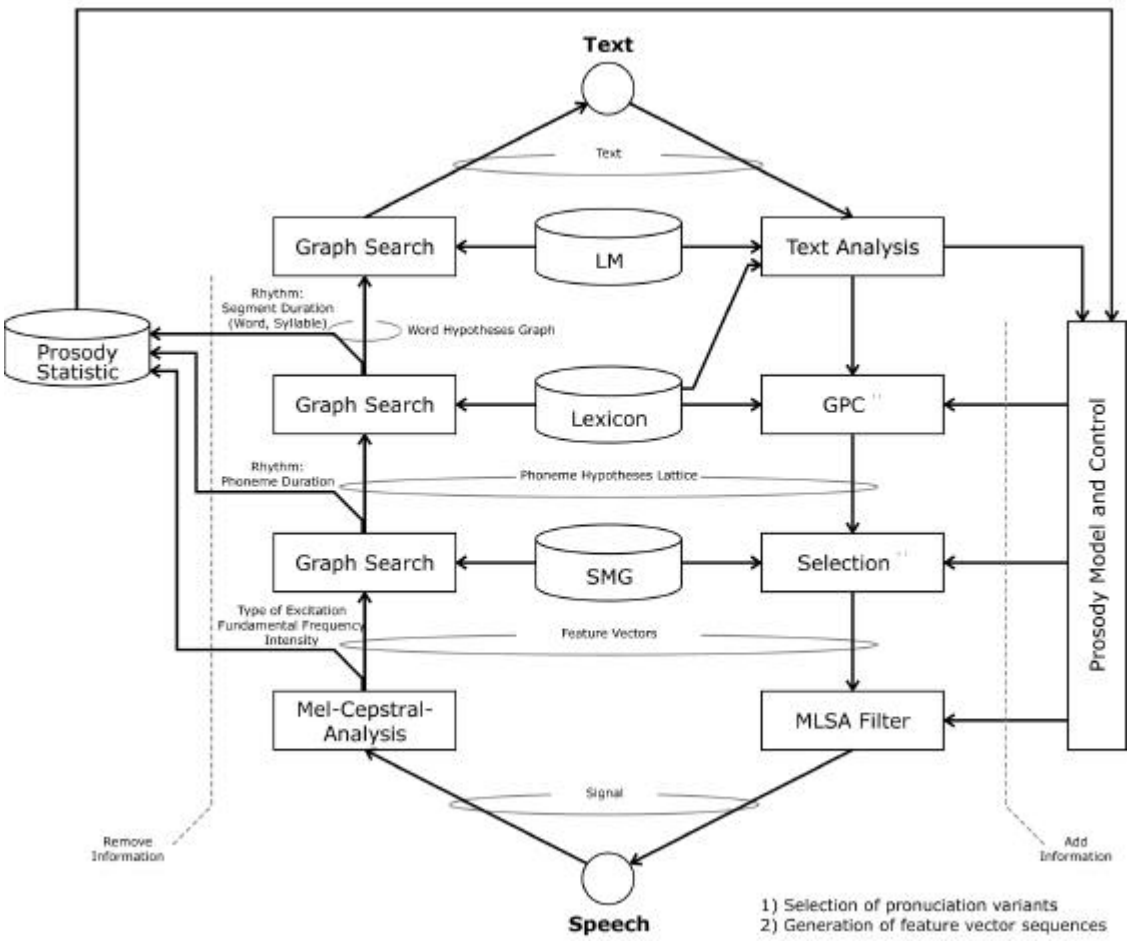

Figure 1: Framework of the integrated system for speech analysis and synthesis.

3.2 Speech Synthesis

The performance of TTS systems seems to approach a stable but insufficient level also. Listeners are judging the overall impression of the quality of synthesized speech around the mean between 0 and the judgment of natural speech [7].

One way to improve the quality of synthetic speech is the corpus based synthesis. This situation is parallel to speech recognition where the application of huge databases leads to the situation that the system has seen every possible phrase during the training phase [8]. In this way, an optimal relation between expense and performance will be obtained neither in recognition nor in synthesis.

We suppose that our integrated approach for synthesis and recognition will be helpful in investigating new algorithms for TTS also. Obviously, the idea to treat the synthesis process as an inverse problem of feature extraction will be useful not only for performance analysis of speech recognizers but also in developing synthesis algorithms. The idea of the so-called HMM synthesis arised at the end of the 1980's (e. g., [9]) and was developed by a Japanese group (summarized in [10]). This approach is basing on a MLSA (mel log spectrum approximation) filter [11] which serves to invert the extraction process of mel-cepstral parameters. A similar method has been applied to the Czech language also [12]. The inversion of other feature extraction algorithms was discussed in the literature, e. g. [13] for the FTT. For our first experiments with the integrated system, the mel-cepstral parameters have been preferred also because they show good results in speech recognition as well.

The implementation of the synthesis component of our system is discussed in more detail in [14]. Below, we want to describe the structure and the training process of the unified databases which we use on the phoneme level utilizing stochastic Markov graphs (SMG) and on the word level utilizing a pronunciation dictionary which is generated by a data driven method.

4. STOCHASTIC MARKOV GRAPHS

We use stochastic markov graphs (SMG) instead HMMs [16] for the acoustic modeling of phones. These models are used in both, speech synthesis and speech recognition. SMGs were first introduced in [15] for the recognition task. The advantage of using SMGs lies in their enhanced capability of modelling trajectories in the

feature space. However, a powerful prediction of feature trajectories is even more important for synthesis. Our trainings procedure resembles the procedure described in [15]. We use 3 state forward connected HMMs with mixtures of four Gaussian distributions as starting point for building the SMGs. We convert an HMM into a SMG by creating a separate node for each base function of the mixture represented by each HMM state. The base functions and graph structures of the resulting models are obtained using the Viterbi training algorithm. In our first experiments we used 3.5 hours of labeled speech for training.

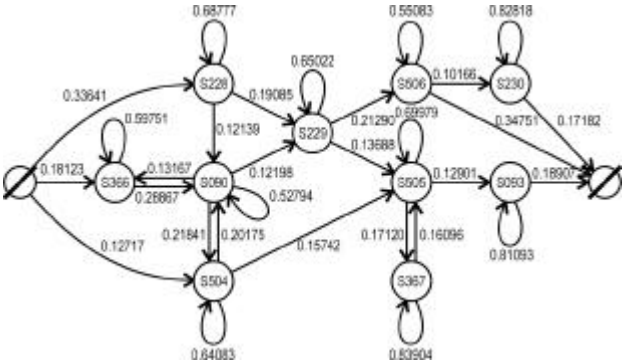

Figure 2: SMG model of phoneme /i:/

During the training not only the distributions of the states are reestimated but also the transition probabilities between SMG states. After the actual training we prune the structure of the SMGs by deleting all transitions that are less probable than a certain threshold and by removing thereby nascended dead paths. Figure 2 shows an example of a trained SMG model for phoneme /i:/ (pruning threshold: 0.07).

To synthesize speech from the SMG models we implemented an algorithm for extraction of the appropriate feature vector sequence depending on a given target phone duration, i.e. the number of feature vectors to be produced. The selection algorithm finds the best path through the model by considering the self-transition probabilities for modeling the state durations within the model. The resulting parameter sequence is synthesized by a MLSA filter [11].

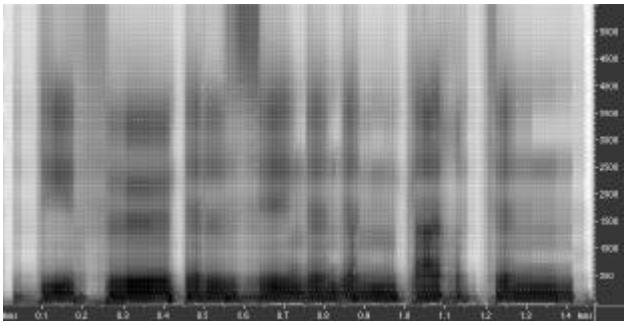

Figure 3: Spectrum of synthesized signal for the German sentence /Die Voegel singen im Garten./

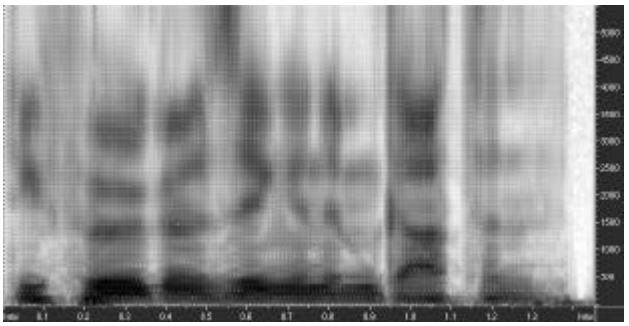

Figure 4: Spectrum of signal resynthesized from MFC parameters.

We obtain the control parameters, like duration and fundamental frequency, from our multilingual TTS system DreSS. Figure 3 shows the spectrum of the synthesized German sentence /Die Voegel singen im Garten./. The original signal, resythesized from MFC parameters, is shown in Figure 4.

Informal tests showed that the quality of synthesized speech using SMGs is slightly better than speech synthesized using HMMs. The resulting speech is intelligible but still lacks naturality of the original signal.

For the recognition task we used a hypotheses based approach. Hypotheses on phones are generated by comparing the feature vector sequence of a speech sample with the parallized SMG model by means of a DP search. The generation of hypotheses is subject to certain restrictions of start and ent time, duration and emission probability (see [18]). The recognition accuracy is calculated by comparing the hypotheses lattices with the original label sequence. An overview of the algorithm is given in Figure 5.

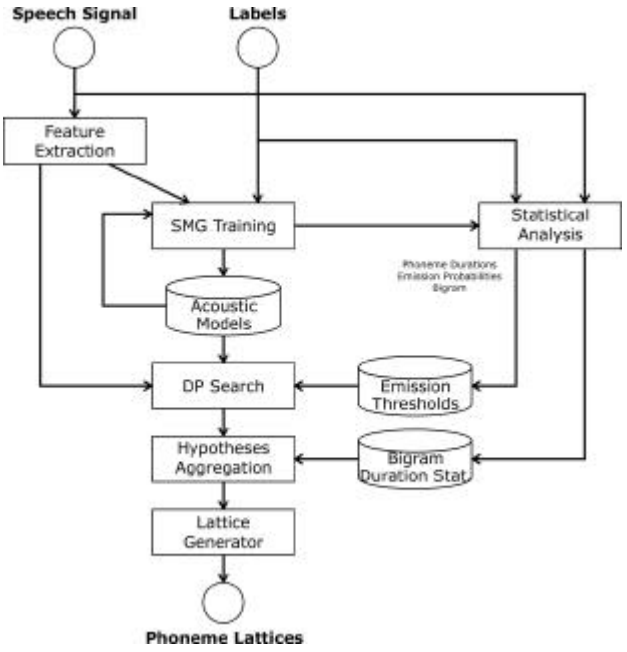

Figure 5: Training of SMGs and application in a hypotheses based phoneme recognizer

To evaluate the suitablility of the chosen Mel-Cepstral-Analysis we trained two sets of SMG models, one using the MFC parameters and one using a 30 channel Mel-filter bank. The speech signal was sampled at 16kHz and windowed by a 25 ms Hamming window with 5 ms shift. The feature vector consists of 63 parameters including energie coefficient, their delta and delta-delta coefficients. After feature extraction the dimension of the feature vector was reduced to 24 parameters by decorrelating the features and omiting axes with minor variance. The experiments showed a similar recognition rate for model trained with MFC parameters and those trained with Mel-filter parameters.

5. THE PRONUNCIATION DICTIONARY

One of the central knowledge sources used by both, the speech recognizer and the speech synthesizer, is the pronunciation lexicon. Our lexicon incorporates pronunciation variants. In most cases the usage of a limited set of pronunciation variants leads to slight improvements of the performance of a speech recognizer (see e.g. [19], [20], [21]). For speech synthesis, pronunciation variants are particularly interesting in view of generating optimal phonetic transcriptions depending on speaking style and speaking rate.

In our unified approach we use a graph based pronunciation lexicon which is automatically learned from speech data. The dictionary training procedure was introduced in [22]. It bases on a phoneme recognizer which generates phoneme hypotheses lattices for the speech samples in the training set. In a second stage, these lattices are compared with a pronunciation model of the speech sample which is assembled from a initial canonical lexicon, or from a previous training iteration, according to the orthographic transcription of the speech sample. Insertions and omissions are explicitly modeled by additional edges in the pronunciation model. Comparison of the phoneme hypotheses lattice and the pronunciation model takes place by a two dimensional DP search which is capable of finding the most probable common path through two graphs. In most cases there will not be an exact match between the two graphs. The deviation between the graphs describe unseen pronunciation variants. The search yields the minimum set of nodes and edges which have to be added to the word models in the lexicon in order to incorporate the new found pronunciation variants. After a training session we post-process the obtained lexicon in order to remove statistically irrelevant variants. A discussion of experiments with this dictionary training approach can be found in [17].

6. CONCLUSION

We introduced an architecture for a unified speech recognition / speech synthesis system. In a first step we implemented stochastic markov graphs as acoustic models. Synthesis experiments as well as recognition experiments proofed the suitability of these models for both tasks. On word level we use a automatically trained variant lexicon for recognition and for the graphem to phoneme conversion in the synthesis task.

Future work will focus on improving the synthesis quality by optimizing the feature selection algorithm and on integrating higher levels of processing such as language and semantic models.

7. REFERENCES

[1] Hoffmann, R., “A multilingual text-to-speech system”, *The Phonetician 80/II, 1999*, 5-10

[2] Hoffmann, R., Ketzmerick, B., Kordon, U., and Kürbis, S., “An interactive tutorial on text-to-speech synthesis from diphones in time domain”, *Proc. Eurospeech 1999, Budapest,* 639 – 642.

[3] Hoffmann, R., and Wolff, M., “Framework design and implementation of web-based tutorials in spoken language engineering”, *Proc. IEEE ICME, New York, July 30 – Aug 2, 2000.*

[4] Lippmann, R. P., “Speech recognition by machines and humans”, *Speech Communication 22 (1997)*, 1 – 16.

[5] Flach, G., Hoffmann, R., and Rudolph, T., “Eine aktuelle Evaluation kommerzieller Diktiersysteme”, *Proc. KONVENS 2000, Ilmenau (Germany),* Oct 9 – 12, 2000.

[6] Chase, L., “Blame assignment for errors made by large vocabulary speech recognizers”, *Proc. Eurospeech 1997, Rhodes (Greece)*, 1563 – 1566.

[7] Hoffmann, R., et al., “Evaluation af a multilingual TTS system with respect to the prosodic quality”, *Proc. Int. Congr. of Phonetic Sciences (ICPhS) 2000, San Francisco*, 2307 – 2310.

[8] Bourlard, H., “Towards increasing speech recognition error rates”, *Proc. Eurospeech 1995, Madrid*, 883 – 894.

[9] Falaschi, A., Giustiniani, M., and Verola, M., “A hidden Markov model approach to speech synthesis”, *Proc. Eurospeech 1989, Paris*, 187 - 190.

[10] Tokuda, K., et al., “Speech parameter generation algorithms for HMM-based speech synthesis”, *Proc. ICASSP 2000, Istanbul*, 1315 - 1318.

[11] Imai, S., Sumita, K., and Furuichi, C., “Mel log spectrum approximation (MLSA) filter for speech synthesis”, *Trans. IECE, vol. J66-A*, 122-129, 1983.

[12] Tychtl, Z., and Psutka, J., “Speech production based on the mel-frequency cepstral coefficients”, *Proc. Eurospeech 1999, Budapest*, vol. 5, 2335 – 2338.

[13] Chalupper, J., and Fastl, H., “Simulation of hearing impairment based on the Fourier time transformation”, *Proc. ICASSP 2000, Istanbul*, 857 - 860.

[14] Eichner, M., Werner, S., Wolff, M., and Hoffmann, R., “Ein kombiniertes Spracherkennungs-/Sprachsynthesesystem auf Phonemebene”, *ESSV, Cottbus (Germany)*, Sep 4 – 6, 2000, *Proceedings = Studientexte zur Sprachkommunikation, vol. 20*.

[15] Wolfertstetter, F., and Ruske, G., “Structured Markov models for speech recognition”, *Proc. ICASSP 1995, Detroit*, 544-547.

[16] Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S., “Speech synthesis using HMMs with dynamic features”, *Proc. ICASSP 1996, Atlanta*, 389-392.

[17] Eichner, M., and Wolff, M., “Data driven generation of pronunciation dictionaries in the German Verbmobil project – discussion of experimental results”, *Proc. ICASSP 2000, Istanbul*, 1687 – 1690.

[18] C.-M. Westendorf, “Erkennung fließender Sprache auf der Basis diskreter Hypothesen – eine Alternative zu HMM?”, *ESSV, Wolfenbüttel (Germany)*, 1995, 85-96

[19] Aubert, X., Dugast, C., “Improved Acousitc Modeling in PHILIPS' dictation system by handling liasons and multiple pronunciations.” *Proc. EUROSPEECH '95*, 767 - 770.

[20] Sloboda, T., Dictionary Learning: “Performance through consistency.” *Proc. ICASSP 1995, Detroit*, 453 - 456.

[21] Lamel, L., Adda, G., “On Designing Pronunciation Lexicons for Large Vocabulary Continuous Speech Recognition.” *Proc. ICSLP 1996*, 6 -9.

[22] Westendorf, C.-M., Jelitto, J., “Learning Pronunciation Dictionary form Speech Data.” *Proc. ICSLP 1996*, 1045 – 1048.