Listen to the Parrot: Demonstrating the Quality of Online Pitch and Formant Extraction Via Feature-Based Resynthesis

Martin Heckmann, Claudius Gläser, Miguel Vaz, Tobias Rodemann, Frank Joublin, Christian Goerick

2008

Preprint:

This is an accepted article published in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). The final authenticated version is available online at: https://doi.org/[DOI not available]

2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Listen to the Parrot: Demonstrating the Quality of Online Pitch and Formant Extraction via Feature-based Resynthesis

Martin Heckmann, Claudius Gläser, Miguel Vaz, Tobias Rodemann, Frank Joublin, Christian Goerick

Abstract— **We present a system for online extraction of the fundamental frequency and the first four formant frequencies from a speech signal. In order to evaluate the performance of the extraction a resynthesis of the speech signal is performed. The resynthesis is based on the extracted frequencies and the energy of the input signal at the formant locations. The extraction of the fundamental frequency and the formants is robust against room echoes and interfering noise. In order to improve the robustness against background noise a noise reduction was implemented. Tests in three rooms of different size at varying distances to the system (up to 8m yielding an SNR of approx. 0 dB) were performed.**

I. Introduction

Due to the large and varying distances between the speaker and the robot the interaction via speech with a humanoid robot like ASIMO is difficult. Room echoes and low speech to noise signal ratios severely impair the signal.

Despite the unfavorable behavior of technical systems, humans perform marvelously well under such conditions [1]. Designing a system based on findings on the functional principles of the human auditory system may lead to a way of overcoming the problems of state of the art systems.

In this paper we present an online system for pitch and formant extraction inspired by results of auditory research. Even tough the focus of the paper is the robust extraction of the parameters we implemented a resynthesis of the speech signal solely based on the extracted parameters in order to asses the quality of the extraction of the parameters. Consequently the system reminds one of a parrot which repeats everything it hears. The three main parts of the system, formant extraction, pitch extraction, and resynthesis (compare Fig. 1), will be detailed in the corresponding sections and be completed by the description of the additional parts in Sec. V. Some examples and comments on the performance will be given in Sec. VI.

II. Formant Extraction

The formant extraction follows the algorithm described in [2]. Before the formants can be tracked first some algorithms to render them more dominant in the spectrogram have to be employed.

Martin Heckmann, Claudius Gläser, Tobias Rodemann, Frank Joublin, Christian Goerick are with the Honda Research Institute Europe GmbH, Carl-Legien-Straße 30, D-63073 Offenbach am Main, Germany {firstname.lastname}@honda-ri.de

Miguel Vaz is with the Department of Industrial Electronics, Universidade do Minho Guimarães, Portugal {mvaz}@dei.uminho.pt

A. Formant enhancement

In a first step the signal is transformed into the spectral domain. Instead of a Fourier transform we use a Gammatone filter bank which models the response of the basilar membrane in the human inner ear and is, therefore, adapted to a biology-inspired system (compare Fig. 1). The signal's sampling frequency is 16 kHz. The filter bank has 100 channels ranging from 80 Hz to 5 kHz.

Next we calculate the amplitude envelope in each frequency channel via rectification and low-pass filtering (compare Fig. 2). On this envelope signal a noise suppression based on Spectral Subtraction implementing a noise level estimation via Minimum Recursive Averaging is applied [3]. Since formants are the resonance frequencies of the vocal tract, their extraction can be improved by eliminating the spectral influence of excitation and radiation contributing to human speech production. It has been shown that this influence can be adequately approximated by a first-order low-pass filter [4] which is valid at least for modal or creaky phonations being by far the most common ones [5]. For this reason, we emphasized the spectral energy by +6 dB/oct.

Additionally, the emphasized spectrogram is smoothed along the frequency axis using a Laplacian kernel adjusted to the logarithmic arrangement of the Gammatone filter banks channel center frequencies. By doing so, the harmonics spread and peaks are formed at formant locations. A subsequent normalization of the filter responses to the maximum at each sample as well as an application of a sigmoidal function further enhances the spectral contrast (compare Fig. 3).

B. Formant tracking

The probabilistic tracking technique we developed is based upon Bayes filters which provide an excellent framework for handling noisy observations [6]. They represent the state at time t by random variables x_t. Thereby uncertainty is introduced by a probabilistic distribution over x_t, called the belief $Bel(x_t) = p(x_t|z_1, \ldots, z_t)$. Their purpose is the sequential estimation of such beliefs over the state space conditioned on all information contained in the sensor data z_t [7].

Let $Bel^-(x_t)$ denote the predicted belief at time t which can be obtained via the application of the formants' underlying dynamics $p(x_t|x_{t-1})$. Then the belief at time t is calculated by correcting the predicted belief according to the preprocessed spectral energy distribution $p(z_t|x_t)$ and a normalization factor α. Thus, the standard Bayesian filter recursion can be written as follows:

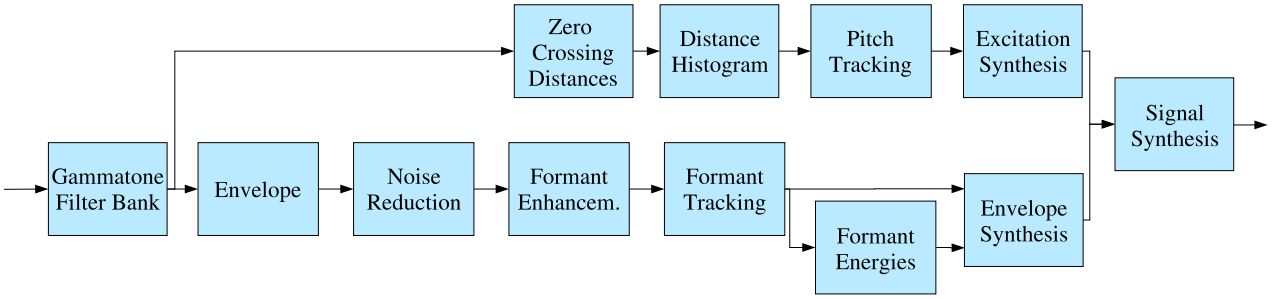

Fig. 1. Overview of the system.

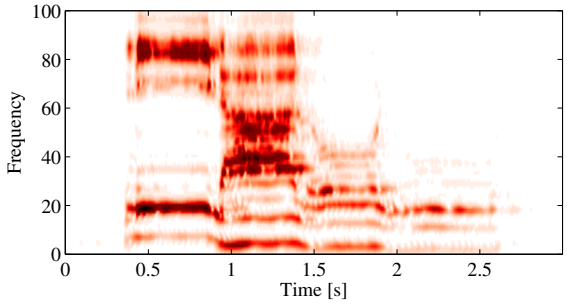

Fig. 2. Envelope of the input signal.

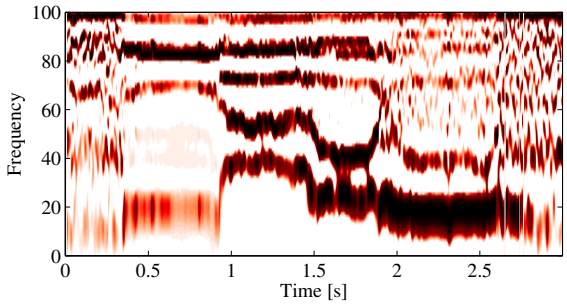

Fig. 3. Result of the formant enhancement.

$$Bel^{-}(x_t) = \int p(x_t|x_{t-1}) \cdot Bel(x_{t-1})\, dx_{t-1} \quad (1)$$

$$Bel(x_t) = \alpha \cdot p(z_t|x_t) \cdot Bel^{-}(x_t) \quad (2)$$

Since standard Bayesian filtering is not an appropriate technique for tracking multiple formants at the same time, we adopted a mixture filtering approach recently introduced in the computer vision community [8]. Thereby, the joint distribution $Bel(x_t)$ is modeled through a non-parametric mixture of M component beliefs $Bel_m(x_t)$ with associated weights $\pi_{m,t}$, so that each target is covered by exactly one mixture component:

$$Bel(x_t) = \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_t) \quad (3)$$

Hence, by substituting the beliefs in Eq. (1) and (2) with Eq. (3) the Bayesian filter recursion can be rewritten with respect to the mixture modeling approach. Furthermore, since we want to estimate formant locations on a discrete grid defined by the channels of the Gammatone filter bank, a grid-based approximation of the belief is chosen. Thus, assuming that the filter bank is composed of N channels, the state space at time t can be written as $X_t = \{x_{1,t}, x_{2,t}, \ldots, x_{N,t}\}$ which leads to the following Bayesian filter recursion:

$$Bel(x_{k,t}) = \sum_{m=1}^{M} \pi_{m,t} \cdot Bel_m(x_{k,t}) \quad (4)$$

$$Bel_m^{-}(x_{k,t}) = \sum_{l=1}^{N} p_m(x_{k,t}|x_{l,t-1}) Bel_m(x_{l,t-1}) \quad (5)$$

$$Bel_m(x_{k,t}) = \frac{p(z_t|x_{k,t}) Bel_m^{-}(x_{k,t})}{\sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_m^{-}(x_{l,t})} \quad (6)$$

$$\pi_{m,t} = \frac{\pi_{m,t-1} \sum_{k=1}^{N} p(z_t|x_{k,t}) Bel_m^{-}(x_{k,t})}{\sum_{n=1}^{M} \pi_{n,t-1} \sum_{l=1}^{N} p(z_t|x_{l,t}) Bel_n^{-}(x_{l,t})} \quad (7)$$

The formulas obtained are quite elegant, since mixture components evolve independently over time. But consequently, belief degenerations (i.e. component distributions becoming more and more diffuse) might occur and cause loosing track of the formants. For this reason, another algorithm which reclusters the beliefs at each time step is needed to ensure the maintenance of multimodality. Assuming such a function exists and returns sets $R_{1,t}, R_{2,t}, \ldots, R_{M,t}$ dividing the frequency range into contiguous formant-specific regions at each time step t, then the belief can be recomputed, so that the mixture approximations of (4) before and after the reclustering procedure are equal in distribution:

$$\pi'_{m,t} = \sum_{x_{k,t} \in R_m} \sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t}) \quad (8)$$

$$Bel'_m(x_{k,t}) = \begin{cases} \frac{\sum_{n=1}^{M} \pi_{n,t} \cdot Bel_n(x_{k,t})}{\pi'_{m,t}} & , \forall x_{k,t} \in R_{m,t} \\ 0 & , \forall x_{k,t} \notin R_{m,t} \end{cases} \quad (9)$$

In this way, previously overlapping beliefs are separated by rearranging their component affiliation depending on associated mixture weights which results in a mixture of consecutive but separated components.

For the necessary segmentation of the frequency range into formant-specific non-overlapping regions we suggested a dynamic programming approach [2]. More precisely, a trellis was built up by which the former problem could be reformulated as the problem of finding the most likely path through the trellis, for which the Viterbi algorithm offers an elegant solution. Since the suggested method relies on the component beliefs at the actual timesteps, the frequency

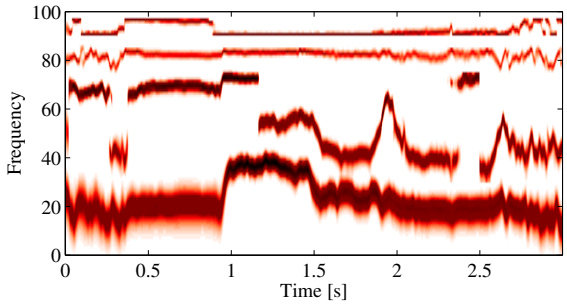

Fig. 4. Result of the Bayesian smoothing.

range is sequentially segmented in an adaptive and computationally efficient manner.

Therewith we are able to apply the Bayesian mixture filtering for tracking the joint distribution of formants while maintaining its multimodality. However, when operating in noisy conditions, a subsequent backward pass on the already obtained filtering distributions $Bel_m(x_{k,t})$ is recommended since it significantly enhances the noise robustness of the algorithm. Bayesian smoothing provides such a mechanism. It aims at recursively estimate a smoothed version $\widehat{Bel}(x_{k,t})$ of the belief, thereby depending on both past and future observations [9]:

$$\widehat{Bel}(x_{k,t}) = p(x_{k,t}|z_1, z_2, \ldots, z_t, \ldots, z_{T-1}, z_T) \quad (10)$$

$$\widehat{Bel}^{-}_m(x_{k,t}) = \sum_{l=1}^{N} \widehat{Bel}_m(x_{l,t+1}) \cdot p_m(x_{l,t+1}|x_{k,t}) \quad (11)$$

$$\widehat{Bel}_m(x_{k,t} = \frac{Bel_m(x_{k,t}) \cdot \widehat{Bel}^{-}_m(x_{k,t})}{\sum_{l=1}^{N} Bel_m(x_{l,t}) \cdot \widehat{Bel}^{-}_m(x_{l,t})} \quad (12)$$

The result of Bayesian filtering followed by a smoothing is visualized in Fig. 4.

The final calculation of exact formant locations $F_m(t)$ can easily be done by picking the peaks of the smoothed component beliefs such that the location of the m-th formant equals the peak location in the smoothed distribution of component m:

$$F_m(t) = \arg\max_{x_{k,t}} \left[\widehat{Bel}_m(x_{k,t})\right] \quad (13)$$

The main advantages of this probabilistic tracking scheme are the estimation of the joint distribution of formants allowing to adaptively resolve ambiguities and the adaptive segmentation of the frequency range into formant-specific regions which takes the interaction of formants into account. Lastly, due to mixture components evolving independently over time, models of the formants underlying dynamics $p_m(x_{k,t}|x_{l,t-1})$ as well as a priori distributions of formant frequencies $p_m(x_{k,0})$ can be chosen for each formant individually. Furthermore, they can be adapted to different conditions such as gender, voicing, or context. Here we used gender-dependent *probability density functions (pdfs)* which can be immediately switched according to the decision of a gender detection system. We assumed that the pdfs can be appropriately modeled by a normal distribution as is shown in Eq. (14) and (15).

$$p_m(x_{k,0}) \propto \mathcal{N}(f(x_k), \mu_m, \sigma_1^{(m)}) \quad (14)$$

$$p_m(x_{k,t}|x_{l,t-1}) \propto \mathcal{N}(f(x_k), f(x_l), \sigma_2^{(m)}) \cdot p_m(x_{k,0}) \quad (15)$$

Here $f(x_k)$ denotes the center frequency of the k-th filter channel. But in contrast to our proposal in [2] we added a mean tendency to $p_m(x_{k,t}|x_{l,t-1})$ which is reasonable since the probability that a formant performs a rising slope is much higher when the formant is actually located at a low than a high frequency. Additionally an enhanced normalization on an extended grid was used for calculating the probabilities. These mechanisms further improved the precision of our method.

III. Pitch Extraction

In our pitch extraction algorithm we combine information residing in the temporal and spectral representation to a more robust algorithm. One part of the algorithm captures the temporal aspects via *Zero Crossing Distances (ZCD)* and the other the spectral aspects via a comb filter (see [10] for more details).

The output of the Gammatone filter bank is the input to the pitch extraction algorithm (compare Fig. 1). For each filter bank channel we calculate the distances between adjacent zero crossings. This distance, more precisely its inverse, codes the frequency of the signal. The zero crossings are similar to the phase locked firing of the neurons in the auditory system, the spike always occurs when the signal rises from negative to positive (if rising zero crossings are used).

A. Zero Crossing Distance Histogram

Partials of a harmonic signal have zero crossings in common. How many zero crossings they share depends directly on their harmonic order relative to the fundamental frequency. For example the first order harmonic shares each second zero crossing with the fundamental. Hence the distance between two zero crossings of the fundamental reappears as the distance between three zero crossings of the first harmonic and so forth. We want to refer to these distances between multiple zero crossings as higher order zero crossing distances.

As a consequence of the reoccurrence of zero crossing distances of the fundamental in the harmonics, a histogram of all distances shows a peak at the fundamental frequency (similar to a so called *all order interspike histogram* of the phase locked firing of the neurons in the auditory system [11]). As not only the distances corresponding to the fundamental frequency but also those of the harmonics reoccur, the histogram shows many spurious side peaks corresponding to the harmonics and sub-harmonics of the true fundamental frequency. Sub-harmonics also occur because, for instance, the second order ZCD of the true fundamental frequency is also the first order distance of the first sub-harmonic ($\frac{1}{2}f_0$).

B. Comb Filter

The activity in the individual channels of the Gammatone filter bank codes the spectral information needed for pattern matching based pitch models. We set up a comb filter for all possible fundamental frequencies with teeth at the harmonics $1 \ldots 7$. The range of possible fundamental frequencies is defined by the resolution of the zero crossing distances and hence by the sampling rate. At a sampling rate of 16 kHz a fundamental frequency of 100 Hz corresponds to 160 samples. The next possible fundamental frequency corresponds to 159 samples, 100.63 Hz respectively. In a scan through all possible fundamental frequencies beginning with the lowest expected fundamental frequency up to the highest one the corresponding comb filters are set up. For each of these comb filters the allocation of the teeth with harmonics of the current fundamental can be checked at each instant in time. The "filter response" of the comb filter is calculated based on the found allocation pattern. The better the found pattern matches the expected pattern, the higher the response.

C. Combining zero crossing distances and comb filtering

In order to determine the allocation of the teeth in the comb filter with harmonics one common way is to use the energy in the band underlying the respective tooth. Here, we deploy the zero crossing distances previously calculated. The Gammatone filter bank has a limited frequency resolution due to a necessary trade off between filter bandwidth and settling time. A decrease in bandwidth and hence an increase in resolution comes at the cost of higher settling time which makes it impossible to analyze transient signals as speech. The ZCDs measure the instantaneous frequency in the time domain and hence are subject to this limitation to a lesser extend.

For each tooth of the comb filter the ZCD with the order corresponding to the harmonic order of the tooth is compared to the ZCD expected for the current fundamental frequency hypothesis. If the deviation between the expected and the measured distance is smaller than a predefined threshold t_Δ the tooth is said to be allocated by the expected harmonic. In the experiments reported later $t_\Delta = 4\%$.

D. Inhibition of Side Peaks

The creation of an allocation table for the comb filters allows to check the found allocation against expected ones. Most of the errors in the histogram are produced at locations at multiples of the true fundamental frequency. When setting up allocation patterns for multiples of the current fundamental frequency hypothesis and checking them against the found allocation pattern, it is possible to inhibit the ones causing the errors (see [10] for more details). Fig. 5 shows the histogram after applying this inhibition mechanism.

E. Pitch Tracking

In order to extract the pitch we apply a tracking algorithm on the final pitch histogram. The pitch tracking algorithm is identical to the formant tracking with the exception that only 1 component is used. After Bayesian filtering and smoothing, the maximum at each sample in time is picked. Finally, the value is converted from a distance measure to frequency.

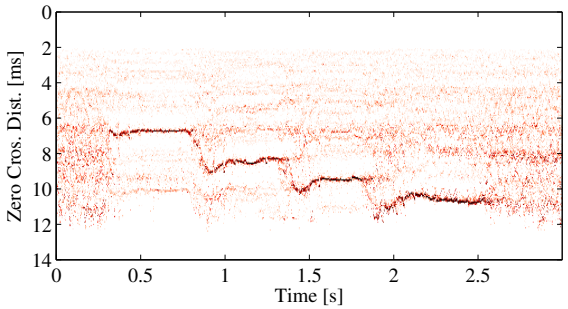

Fig. 5. Zero crossing distance histogram with inhibition of side peaks.

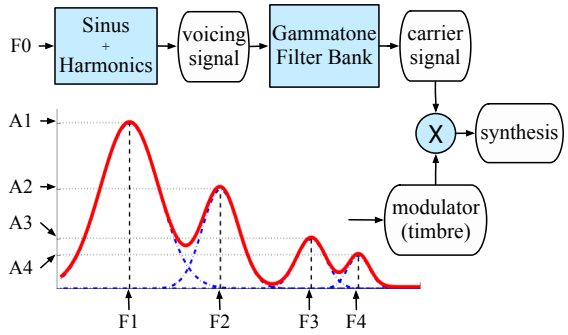

Fig. 6. Resynthesized formants.

IV. Resynthesis

Based on the extracted 9 parameters, the fundamental frequency and the first four formants including their energy, we resynthesize the original speech signal. For this purpose we use a technique similar to a classic channel vocoder [12] (compare Fig. 6). A channel vocoder assumes that speech can be split in a source signal and a time varying filter. This is an abstraction of the human vocal folds and the vocal tract. In our case the source signal is a sinusoid plus corresponding harmonics, with fundamental frequency equal to that extracted from the input speech. It is exclusively voiced in nature, since there is no extraction of frication parameters.

Unlike the classic channel vocoder, the modulator (filter) is derived only indirectly from the input speech. The channels' values are represented via a mixture of four Gaussians, each centered at one of the extracted formant frequencies. The height of each of these Gaussians is borrowed from a representation very similar to that used for the formant enhancement. The sole difference is the use of Gaussian kernels for the smoothing along the frequency axis instead of Mexican hat shaped filters in the formant case. The reason is that the Mexican hat filters enhance the separation between the different formants whereas here we only want to extract the energy in each frequency bin independent on the fundamental frequency. The Gaussians' heights are consequently determined by sampling this final representation at the positions of the formants. The width of each Gaussian is

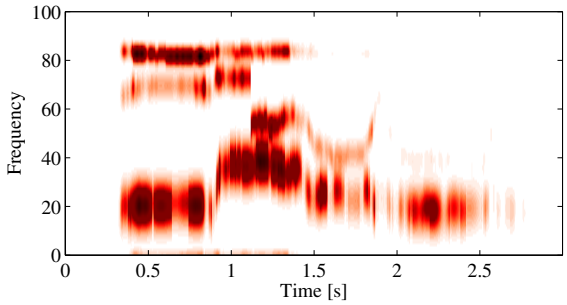

Fig. 7. Resynthesized formants.

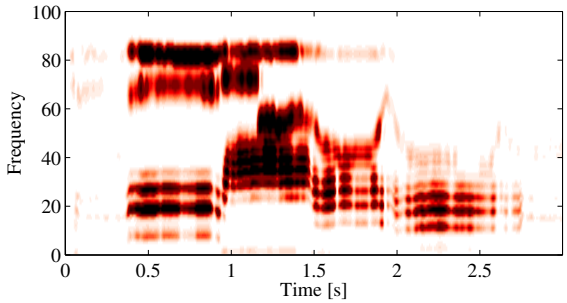

Fig. 8. Envelope of the result of the resynthesis.

predefined. The resulting spectrogram can be seen in Fig. 7.

The final step is to combine the harmonic source signal and the spectral envelope represented by the Gaussians at the formant locations. Therefore, the source signal goes through a channel decomposition via a Gammatone filter bank identical to the one used for the analysis of the input signal. As a consequence, the representation of the source signal and the envelope are identical and can be combined via a simple channel-wise multiplication. In Fig. 8 the final envelope of the resynthesized signal is depicted.

V. Online Integration

For the implementation we used our software design, execution, and monitoring framework for real-time applications *ToolBOS* [13]. This software infrastructure allows us to design applications in a modular way and flexibly distribute them on multiple computers.

The system runs on one computer with an Intel Quad Core processor (Q6600 @ 2.4 GHz). A second, identical computer is used for the visualization of the internal states of the system (e.g. pitch and formant tracks). The speech signal is acquired via a DPA 4060-BM lavalier microphone which is mounted inside of a silicon pinnae with human like shape on a replica of Asimo's head (compare Fig.9).

VI. Results

It is difficult to assess the quality of an online system. In order to make the scenario realistic it is advisable to speak to the system instead of using prerecorded sentences. However, in this case there is no ground truth available and hence the correctness of the extraction can not be judged easily. We previously evaluated the performance of the extraction

TABLE I

Comparison between MFCCs and formants on the male and female part of TIDigits (in percent absolute Word Error Rates).

	clean	babble 6 dB	car 6 dB	white 6 dB
MFCC	0.5	43.8	4.4	75.9
Formants	11.1	37.1	27.9	35.6

algorithms individually. In [10] we showed that our pitch extraction algorithm is very robust against background noise and that the resulting histograms contain substantially less noise than those produced via the autocorrelation. We have not tested our tracking by itself but we demonstrated that the combination of our formant enhancement and our tracking leads to significantly better results on clean [2] and noisy data [14] than state of the art algorithms.

In an additional offline test we trained an HMM with the formant tracks extracted by our system on the subset of TIDigits containing only male and female speakers (no children) [15]. This yielded 8623 utterances in the training set and 8700 utterances in the test set. The HMMs were modeled with HTK and the parameter settings were identical to the Aurora-2 framework [16] (apart from using 16 kHz instead of 8 kHz). We also calculated deltas and double deltas for the formants and the MFCCs. In addition to that we also added the log energy to the formant tracks to have some kind of energy information. As can be seen from Tab. I the recognition based on the formants alone is far inferior to using MFCC features on clean speech. However, when adding noise the recognition based on the formants catches up. Only for car noise, a noise type the MFCCs cope very well with, they remain clearly superior.

In order to evaluate the extraction performance of the complete system we use the intelligibility of the resynthesized speech signal. Errors in the extraction will lead to the generation of unnatural sounds or deviating pitch trajectories. The results can be evaluated via the accompanying video where we talk to the system in two different scenarios. First, close to the microphone and then at a distance of about 8 m (limited by the room size). The speech signal of the speaker in the video is the one captured by the system. In Fig. 9 an image from the video is depicted. As can be seen the intelligibility is very good in the case where we talk close to the microphone and only drops a little bit when talking from far. When judging the intelligibility of the resynthesized signal it has to be taken into account that we do not model unvoiced parts of speech and rather resynthesize all segments as voiced.

So far we tested the system in 3 different rooms with different echo constants τ_{60}:

- $\tau_{60} \approx 625$ ms, size $\approx$ 3 x 5 x 3 m
- $\tau_{60} \approx 810$ ms, size $\approx$ 12 x 11 x 2.8 m
- $\tau_{60} \approx 975$ ms, size $\approx$ 12 x 10 x 3.1 m .

In all rooms we were talking close as well as far from the microphone. Due to the additional noise sources present (computers, air conditions, beamer) we had SNR levels of

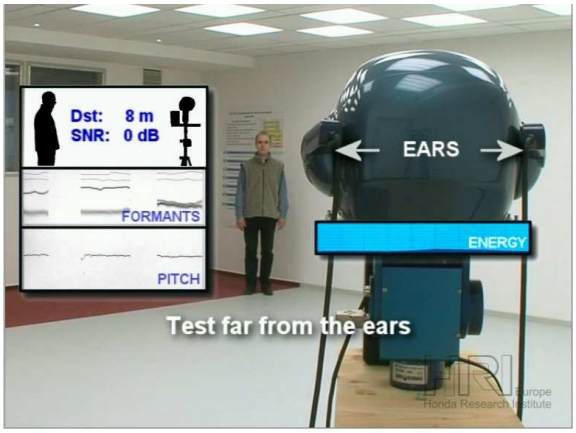

Fig. 9. Test of the system in a room with $\tau_{60} \approx 810\,\mathrm{ms}$ at a distance of 8 m and an SNR of 0 dB.

$\approx 15 \ldots 0$ dB. The SNR levels were estimated based on recordings of the stationary noise signal and the speech signal plus noise. In all cases tested the intelligibility was good. The second scenario in the accompanying video is the most difficult scenario we tested (long distance to the microphone and rather low SNR), consequently the other results were as good or better as this scenario.

The comparison of the spectra of the original speech signal (compare Fig. 2) and the result of the resynthesis (compare Fig. 8) also gives a hint on the extraction quality. As can be seen the important aspects of the original signal are kept despite the fact that only 9 parameters were used for the resynthesis and frication is not modeled. The signal in Fig. 2 was recorded in the smallest room with a distance of $\approx 1\,\mathrm{m}$ to the microphone.

VII. Discussion

We presented an online system which integrates pitch and formant extraction. Evaluation of the system based on visual inspection of the extracted tracks and resynthesis of the original speech signal based on the 9 extracted parameters (f_0, $F_1 \ldots F_4$ and corresponding energies) revealed that it is very robust against changes in the distance between microphone and speaker and the noise level. Even at a distance of 8 m and an SNR of 0 dB the resynthesized signal was only slightly degraded in comparison to a similar signal uttered directly in front of the system. In our view this robustness emerges from the combination of the robust pitch and formant extraction (compare [10], [2]), the efficient Bayesian tracking algorithm and the adaptive noise reduction which is very efficient for the noise encountered in our setup, namely office environments with more or less constant fan noise.

The previously performed offline test on the pitch and formant tracking confirmed these results. However, an additional offline speech recognition experiment revealed that the formant tracks do not carry sufficient information for speech recognition. MFCC features performed significantly better. Only for high noise levels the formant tracks showed better robustness. In our view this poor performance of the formant tracks is to a large extend due to the coarse modelling of plosives and fricatives and the insensitivity to voicing. In contrast to the used HMM model humans seem to be able to compensate for this information loss as the resynthesized sentences are also well comprehensible under difficult conditions. However, the robustness of our features in the presence of noise proved superior to that of MFCCs. Therefore, we are confident that the features based on formant tracks can also obtain superior recognition performance for low noise levels when a better modeling of plosives and voicing is included.

VIII. Acknowledgment

We want to thank Christophe Lorin for his help in implementing important parts of the system, Marcus Stein and Benjamin Dittes for their support with ToolBOS related problems, and last but not least Mark Dunn for his support with the computing hardware and the audio drivers.

References

[1] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, Nov. 08 1997.

[2] C. Gläser, M. Heckmann, F. Joublin, C. Goerick, and H.-M. Groß, "Joint estimation of formant trajectories via spectro-temporal smoothing and bayesian techniques," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, Honolulu, Hawaii, 2007, pp. 477–480, IEEE.

[3] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin, and C. Goerick, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc IEEE/RSJ Int. Conf. on Robots and Intell. Syst. (IROS)*, Beijing, 2006, IEEE Press.

[4] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, New York, 2nd edition, 2000.

[5] DG Childers and CK Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoust. Soc. of America*, vol. 90, pp. 2394, 1991.

[6] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[7] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian filters for location estimation," *Pervasive Comput.*, vol. 2, no. 3, pp. 24–33, 2003.

[8] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multimodality through mixture tracking," in *Proc. IEEE Int. Conf. Comp. Vision (ICCV)*, 2003, vol. 2, pp. 1110–1116.

[9] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for nonlinear time series," *J. of the American Stat. Assoc.*, vol. 99, no. 465, pp. 156–168, 2004.

[10] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, ISCA.

[11] P. A. Cariani, "Temporal codes and computations for sensory representation and scene analysis," *IEEE Trans. Neural Networks*, vol. 15, pp. 1100–1111, 2004.

[12] B. Gold and N. Morgan, *Speech and audio signal processing*, John Wiley & Sons, Inc. New York, 2000.

[13] Antonello Ceravola, Marcus Stein, and Christian Goerick, "Researching and developing a real-time infrastructure for intelligent systems – evolution of an integrated approach," *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 14–28, Jan. 2008.

[14] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Auditory-based formant estimation in noise using a probabilistic framework," in *Proc. INTERSPEECH 2008*, Brisbane, Australia, 2008, ISCA.

[15] R. Leonard, "A Database for Speaker-independent Digit Recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 1984, vol. 9.

[16] D. Pearce and H.G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Int. Conf. on Spoken Lang. Proc.* 2000, ISCA.