AMISCO: The Austrian German Multi-Sensor Corpus

Hannes Pessentheiner, Thomas Pichler, Martin Hagmüller
Signal Processing and Speech Communication Laboratory
Graz University of Technology, Inffeldgasse 16c/EG, 8010 Graz, Austria
hannes.pessentheiner@tugraz.at, thomas.pichler@student.tugraz.at, hagmueller@tugraz.at

Abstract

We introduce a unique, comprehensive Austrian German multi-sensor corpus with moving and non-moving speakers to facilitate the evaluation of estimators and detectors that jointly detect a speaker's spatial and temporal parameters. The corpus is suitable for various machine learning and signal processing tasks, linguistic studies, and studies related to a speaker's fundamental frequency (due to recorded glottograms). Available corpora are limited to (synthetically generated/spatialized) speech data or recordings of musical instruments that lack moving speakers, glottograms, and/or multi-channel distant speech recordings. That is why we recorded 24 spatially non-moving and moving speakers, balanced male and female, to set up a two-room and 43-channel Austrian German multi-sensor speech corpus. It contains 8.2 hours of read speech based on phonetically balanced sentences, commands, and digits. The orthographic transcriptions include around 53,000 word tokens and 2,070 word types. Special features of this corpus are the laryngograph recordings (representing glottograms required to detect a speaker's instantaneous fundamental frequency and pitch), corresponding clean-speech recordings, and spatial information and video data provided by four Kinects and a camera.

Keywords: Joint detection, joint estimation, direction of arrival, fundamental frequency, source localization, position-tracking, pitch-tracking, multi-channel, multi-room, multi-modal, glottogram, voice-controlled environment, Kinect, MEMS.

1. Introduction

In computational auditory (Bregman, 1990; Wang and Brown, 2006) and acoustic scene (Teutsch, 2007) analysis, signal parameters often need to be associated with their origin. For instance, to describe an acoustic scene, we need to detect and localize events, separate them from each other, characterize and interpret them (Imoto et al., 2013; Kwon et al., 2009; de Cheveigné and Slama, 2006). To localize and characterize such events, we jointly detect or estimate each source's parameters to avoid data association performed by additional algorithms.

1.1. Problem Definition

In the area of distant speech enhancement (Vary and Martin, 2006; Woelfel and McDonough, 2009), several research teams dedicated their time to jointly detect or estimate a source's direction of arrival (DOA) and fundamental frequency (f_0) with two or more microphones. Finding these parameters is a prerequisite to improve, e.g., the word accuracy rate of a speech recognizer by applying beamforming or source separation algorithms. They applied their algorithms to synthesized harmonic signals (Jensen et al., 2010; Kronvall et al., 2014), signals from musical instruments (Jensen et al., 2013; Jensen et al., 2015), certain parts of filtered clean-speech signals (Ngan et al., 2003; Karimian-Azari et al., 2013), synthetically spatialized signals (Karimian-Azari et al., 2013), or speech signals without having a reliable ground truth of the f_0 (Habib and Romsdorfer, 2013). One thing they all had in common was no access to multi-channel speech data recorded in real environments and labeled with f_0s and DOAs or positions in space, because such data did not exist.

1.2. Problem Solution

To make such data publicly available, we, therefore, set up a corpus containing Austrian German multi-sensor, multi-channel speech recordings in a real environment, shown in Fig. 1 and Fig. 2, labeled with a speaker's f_0s, positions, orientations, and other parameters. The new corpus offers glottograms that can be used in prosody analysis, speech coding, speaker identification, as well as speech recognition. They are a prerequisite to evaluate pitch-detectors, e.g., YIN (de Cheveigné and Kawahara, 2002) and RAPT (Talkin, 1995), and pitch-trackers (Wohlmayr et

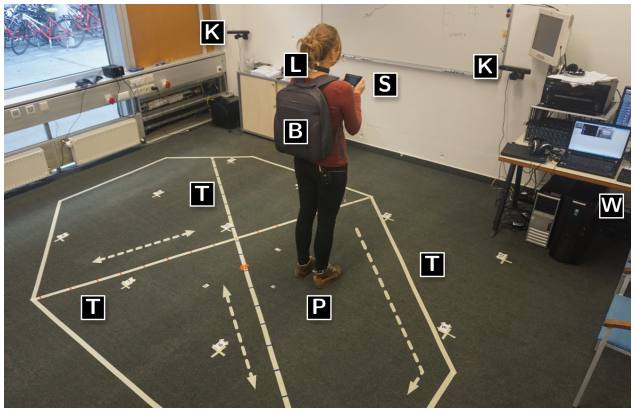


Figure 1: A speaker reads sentences shown on the smartphone's screen (S) at position ten (P) facing east in our meeting room. There are two Kinects (K) next to the whiteboard, a work station (W) on the right-hand side, and marked trajectories (T) and positions (P) on the floor. The speaker is wearing a back bag (B) containing a battery-driven laryngograph and transmitters. The laryngograph's sensor (L) is mounted on the speaker's neck. The bright arrows on the floor mark the directions of movement, the small cross-shaped markers represent the positions with four orientations. The red spot on the left-hand side of the speaker marks the pentagonal array's center (mounted on the ceiling).

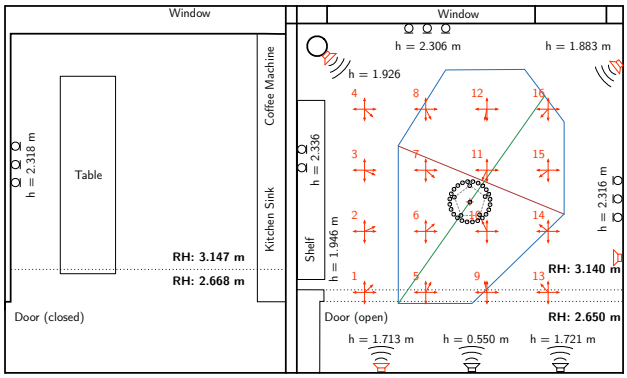


Figure 2: Floorplan of our recording-environment. The kitchen features three, the meeting room 38 microphones. We equipped the meeting room with a video camera and four Kinects (highlighted as red chassis: top right, center right, bottom left, top left). There were five loudspeakers (illustrated as chassis: bottom left to bottom right, top left, top right) mounted on the walls. The red markers on the floor represent the positions and orientations, the blue, green, and red lines the trajectories.

al., 2011). Furthermore, the corpus is suitable for linguistic studies, various machine learning and (multi-modal and multi-channel) signal processing tasks, and studies related to a speaker's fundamental frequency.

1.3. Related Work

There are outstanding corpora available; however, they do not meet our requirements to jointly detect and estimate DOAs and f_0s.

On the one hand, there are three corpora that include glottograms and/or laryngograph recordings: the Mocha-TIMIT database (Wrench, 2000), the Keele corpus (Plante et al., 1995), and the PTDB-TUG corpus (Pirker et al., 2011). The Mocha-TIMIT (Wrench, 2000) includes recordings from a male and a female speaker sampled at 16 kHz only, which is usually not high enough for automatic speech recognition. The PTDB-TUG (Pirker et al., 2011) and the Keele corpus (Plante et al., 1995) contain single-channel and close-talking speech recordings, which cannot be used for multi-channel experiments.

On the other hand, there are corpora containing multi-room and multi-channel recordings: the ATHENA corpus (Tsiami et al., 2014), the DIRHA-GRID corpus (Matassoni et al., 2014), and the GRASS corpus (Schuppler et al., 2014a). Unfortunately, the provided glottograms and laryngograph recordings are highly distorted (in the latter case) or missing (in the former cases).

Moreover, none of these corpora contains moving speakers, and none of them but (Schuppler et al., 2014a) contains any (Austrian) German speech recordings, which are indispensable for experiments with (Austrian) German voice-controlled systems. Almost all corpora mentioned before lack video recordings, which are important for audio-visual experiments and quality assurances. In (Le Roux et al., 2015) they reported that the perfect data set is out of reach when they perform automatic speech recognition using microphone arrays. All these lacks led us to set up a new, unique, and comprehensive corpus called "AMISCO: The Austrian German Multi-Sensor Corpus."

2. Data Collection and Data Editing

2.1. Speakers

The corpus contains read-speech produced by 24 speakers, balanced male and female, aged between 25 and 52, 23 of them with Austrian German, and one with German German to be able to draw a rough comparison between both variants' pronunciation. Since discussing the differences between Austrian German and German German is out of this article's scope, we refer to (Moosmüller et al., 2015; Schuppler et al., 2014a) for more information about this topic. All speakers but one were born in Austria; one was born in Germany. Those who were born in Austria grew up in non-western provinces, which ensures reduced dialectal variations. At the time of recording, all of them lived in Graz, Austria, and exhibited a higher education with at least a university degree. We asked each speaker to fill in a short questionnaire to get an overview of his/her language skills, education, diseases of the vocal tract, and other personal parameters (e.g., body height which is a prerequisite to approximate the position of the signal-emitting head in space).

2.2. Equipment

To guarantee high-quality recordings, we employed a high-end sound card with a word clock module to connect the computer with three pre-amplifiers and three audio interfaces. We sampled the wave field by employing different types of arrays: a wall-mounted 2-element linear array with a length of 30 cm, three wall-mounted 3-element linear arrays with a length of 60 cm, and a ceiling-mounted 5-element pentagonal array with a diameter of 54.44 cm and a microphone in its center. These microphone arrays featured omnidirectional boundary microphones. We also used MEMS omnidirectional microphones connected to three microcontroller-boards to set up a circular array with a diameter of 61.90 cm; it surrounded the pentagonal array. The microcontroller-boards were connected to the server via USB. To facilitate close-talking and laryngograph recordings, we employed a headset microphone, a portable laryngograph, and two wireless transmitters (of a professional digital wireless microphone system) connected to these devices. The transmitters and the receivers were synchronized with the central word clock. Additionally, we used four Kinects (featuring the Microsoft Kinect skeleton tracker based on SDK v1.8) and a video camera; they captured 30 frames per second. A multi-core computer operated the Kinects and transmitted the captured data via TCP/IP to our main server. For recording and post-processing, particularly for synchronizing audio with video data, we employed digital audio workstations. In total, we applied 43 acoustic sensors (including the laryngograph), four Kinects, and a video camera.

2.3. Recording Environment

We did the recordings in rooms that are characteristic for active or ambient assited living and staff meetings (DIRHA.fbk.eu, 2015; Hagmüller et al., 2015), i.e.,

a kitchen at home and a meeting room in a company. In these rooms, distant speech enhancement, e.g., localization and characterization of multiple sources (Pessentheiner et al., 2015), has to be applied for successful speech recognition. Figures 1 and 2 show our recording environment consisting of a meeting room with dimensions $(5.3 \times 5.8 \times 3.1)$ m and a kitchen with dimensions $(4.0 \times 5.7 \times 3.1)$ m. The reverberation time in the meeting room is around $T_{60,c} \approx 500$ ms, whereas the reverberation time in the kitchen is about $T_{60,k} \approx 700$ ms when placing signal-emitting sources in the meeting room.

2.4. Calibration

To guarantee a well-calibrated recording system, we generated a diffuse noise-field, measured the averaged captured power (over all frequencies) of each channel, and adjusted each channel's gain to obtain the same averaged captured power for each channel. We employed five preinstalled hi-fi loudspeakers in the meeting room to play back white noise. To ensure a diffuse noise-field, we measured the A-weighted equivalent sound pressure level in front of each microphone by applying a sound level meter. Measurements revealed level-differences between ± 1 dB. To calibrate the Kinects, we selected four position-markers in the center of the meeting room (at position 6, 7, 10, and 11), which were in the visual field of all Kinects. We computed the markers' coordinates and walked within the area spanned by those markers on well-defined paths to determine the deviations. Then, we adjusted the Kinects' positions and orientations by hand and measured both parameters by applying a laser distance meter. After repeating the measurement procedure several times, we averaged all measurements and entered the resulting coordinates in the Kinects' config-files to improve their accuracy.

2.5. Recording Procedure

Each speaker read items that appeared on a smart-phone's screen. At the beginning, we informed the speaker about the purpose of the recording, and he/she signed a statement of agreement (e.g., to ensure that we preserve the speaker's anonymity). Afterwards, we instructed the speaker about the overall procedure and equipped him/her with a headset, a back bag containing a battery-driven laryngograph and wireless transmitters. On the neck close to the larynx we mounted the laryngograph's sensors. We recorded the speaker in one session (50-60 min) composed of three sub-sessions (10-12 min) and short breaks where we informed the speaker about the upcoming tasks. In sub-session 1, the speaker read 104 short items at positions, which were selected uniformly at random, with 5 different orientations per position: north, east, south, west, and center of the pentagonal array. In sub-session 2, the speaker walked at constant speed along predefined trajectories marked at the floor. We split this sub-session into three parts: (1) reading 24 long sentences and walking along the heptagon-shaped trajectory clockwise, walking along the inner, straight trajectories from (2) west to east and (3) north to south, and vice versa, and reading in total 40 long sentences. Sub-session 3 was identical to sub-session 1, except that the speaker read 64 long items. During the whole session, two assistants (they were sitting in the same room) supervised the speaker by verifying the read items, the positions and orientations, and the speaker's gait velocity in sub-session 2. Furthermore, they triggered the change of sentences-to-read manually. In case of any bloopers, interferences, or other problems, they told the speaker to stop, to wait, and to read the item again.

	Min-SNR [dB]	Max-SNR [dB]	Avg-SNR [dB]
Headset	23.58	52.23	38.67
CPR 1-2	18.97	36.43	24.73
CPR 3-5	19.14	35.75	24.84
CPR 6-8	19.32	36.66	25.09
CPR 9-14	19.06	37.14	24.33
Kitchen 15-17	17.27	31.54	21.35
MEMS M1	20.68	42.14	27.46
MEMS M2	21.97	45.09	29.62
MEMS M3	21.50	44.53	29.15

Table 1: Minimum, maximum, and average signal-to-noise ratio over all speakers in dB for one microphone of each microphone array and the headset. CPR denotes the meeting room.

2.6. Acquisition Data

The recorded utterances represent read speech, phonetically balanced sentences, commands, and digits. The sentences were identical to those used in the GRASS corpus (Schuppler et al., 2014b), and the orthographic transcriptions include around 53,000 word tokens and 2,070 unique word types. We recorded the utterances with a sampling frequency of 48 kHz and encoded them with PCM S24 LE (araw). After synchronizing the audio recordings with the video data we set markers ~ 0.5 s before and after each utterance by hand. Then, we exported the markers as a text file (one file per session) and split the original multi-channel files and the Kinects' skeleton tracks into chunks. Table 1 shows the minimum, maximum, and average signal-to-noise ratio (SNR) over all speakers for one microphone of each microphone array and the headset in dB. Apart from speech recordings (see Fig. 3 and Fig. 4), we provide estimated f_0s, glottograms, positions and orientations of each speaker, files containing additional information about the speaker and the scenarios, as well as orthographic transcriptions. Moreover, we provide different noise recordings in the meeting room and kitchen: moving chairs, smashing and closing doors, running water-tap, etc.

2.7. Post-Processing

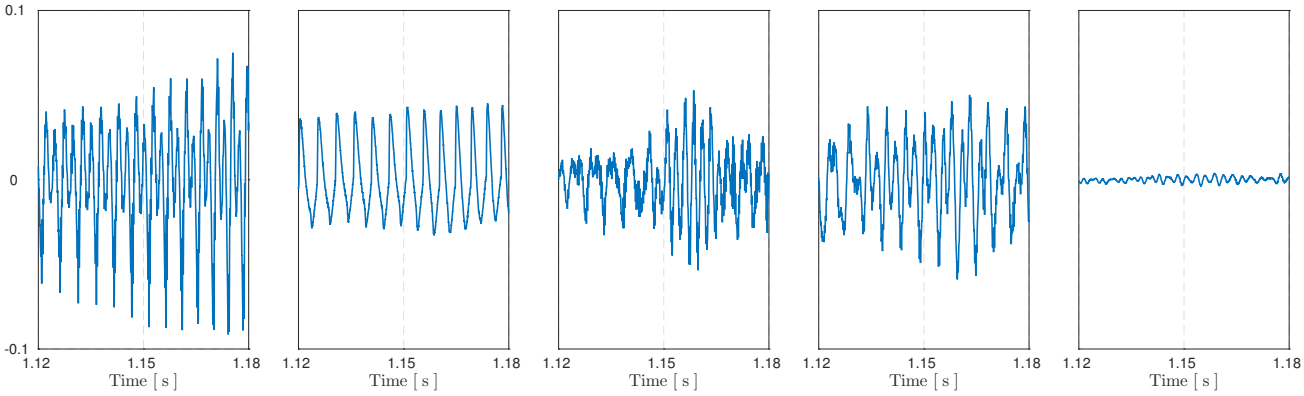
2.7.1. Signal-to-Noise Ratio

We determined segmental SNRs of each speaker's recorded utterance by applying a short-term power estimation utilizing a first-order IIR smoothing of the signal's instantaneous power (Hänsler and Schmidt, 2004) of a randomly selected microphone of each array according to

$$P_s[n] = (1 - \gamma[n]) \cdot x^2[n] + \gamma[n] \cdot P_s[n-1]$$

with

$$\gamma[n] = \begin{cases} \gamma_r & \text{if } x^2[n] > P_s[n-1] \\ \gamma_f & \text{otherwise} \end{cases}.$$



(a) Headset. (b) Glottogram. (c) CPR 1. (d) MEMS M1/1. (e) Kitchen 15.

Figure 4: Time signals of a section (60 ms) of the (German-language) sentence [jeː neːa dɛɐ tsaɪgɐ aʊf axt kaːm dɛstɔ ʊnruːɪgɐ vʊɐdən diː lɔʏtə] (IPA) read by speaker two. To plot these figures, we used the file named `02_f_long_2_<LABEL>_028_6_2.wav`, where `<LABEL>` is a wildcard. For instance, the used audiofile in plot (c) is `02_f_long_2_cpr1-2_028_6_2.wav`, where `02` denotes the speaker's id-number, `f` is the gender, `long_2` is the session, `cpr1-2` represents the room's label (cpr) and the used microphones (1 and 2), `028` denotes the id-number of the spoken item, `6` is the position, and The figures show the signals of (a) the headset microphone, (b) the laryngograph (represented as a glottogram), (c) the first microphone of the linear array labeled as CPR1-2, (d) a MEMS microphone of the circular array's first MEMS module, and (e) the first microphone of the kitchen's linear array. In comparison to (a), the signals in (c-e) are time-shifted and filtered due to the time-differences of arrival and the influence of the room.

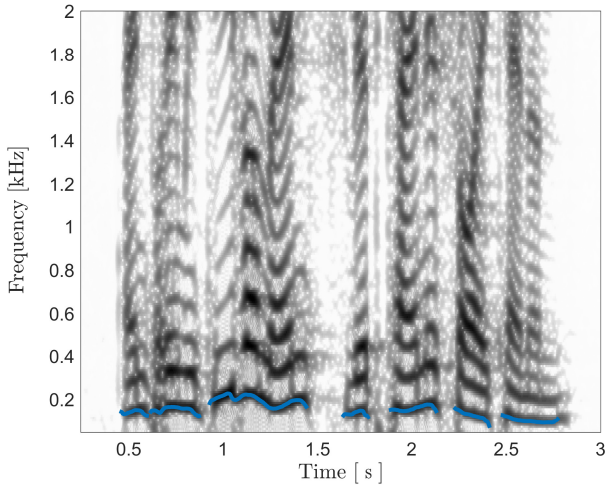


Figure 3: Spectrogram of a male speaker's speech signal recorded with a headset. The (German-language) utterance is [aɪm priːmiːtiːvɐ mɛnʃ vɪrt kaɪmə ʃɔʏ kɛnnən] (IPA). The f_0's ground truth values are marked with a blue line. The corresponding audiofile used for this figure is `22_m_short_1_wireless_042_9_1.wav`, where `22` denotes the speaker's id-number, `m` is the gender, `short_1` is the session, `wireless` denotes the headset's recording, `042` is the id-number of the spoken item, `9` is the position, and `1` is the orientation, respectively.

Variables γ_r and γ_f are smoothing constants for rising and falling signal edges, $x[n]$ is the input signal at index n, and $P_s[n]$ is the smoothed instantaneous power of the signal. Then, we estimated the local background noise power as follows:

$$P_b[n] = (1+\epsilon) \cdot \min\left(P_s[n], P_b[n-1]\right),$$

where ϵ is a small positive constant, which controls the maximum speed for increasing the estimated noise level. After this, we computed the power ratio,

$$P[n] = 10 \cdot \log_{10}\left(\{P_s[n] - P_b[n]\}/P_b[n]\right),$$

and averaged all values of $P[n]$ above a certain threshold μ yielding the average SNR per audio file:

$$\mathrm{SNR}^{(u)} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} P[k]$$

with $\mathcal{K} = \{n \mid \forall n : P[n] \geq \mu\}$, where $|\mathcal{K}|$ is the cardinality of set $\mathcal{K}$. After averaging the SNR of each speaker's utterance, we obtain the overall SNR per microphone:

$$\mathrm{SNR} = \frac{1}{N_u} \sum_{u=1}^{N_u} \mathrm{SNR}^{(u)},$$

where N_u is the number of all utterances. We chose $\gamma_r = 0.99$, $\gamma_f = 0.97$, $\epsilon = 2 \cdot 10^{-5}$, $\mu = 15$ dB, and $P_b[0] = P_s[0] = x[0]^2$ as initial values.

2.7.2. Resampling & Filtering Skeleton Tracks

Since the Kinects delivered unequally spaced detections in time, the data points had to be resampled with equally spaced 30 fps. Knowing that the speakers had a constant gait velocity, we resampled the resulting skeleton tracks by considering linear interpolation, which yielded data points with equally spaced time-intervals. The measurement of the Kinects' positions with a laser distance meter by hand

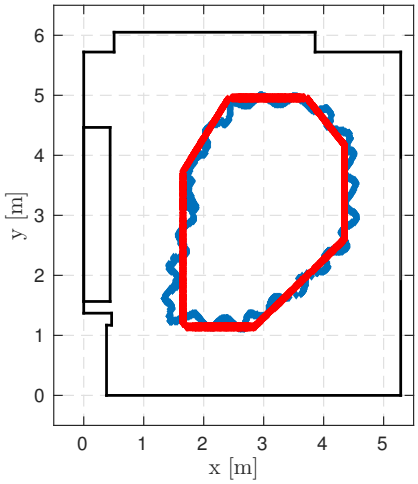


Figure 5: The original (blue and curvy) and modified (red and straight) skeleton tracks represented as trajectories in our meeting room's floor plan. These trajectories correspond to the movements of speakers in sub-session 2.

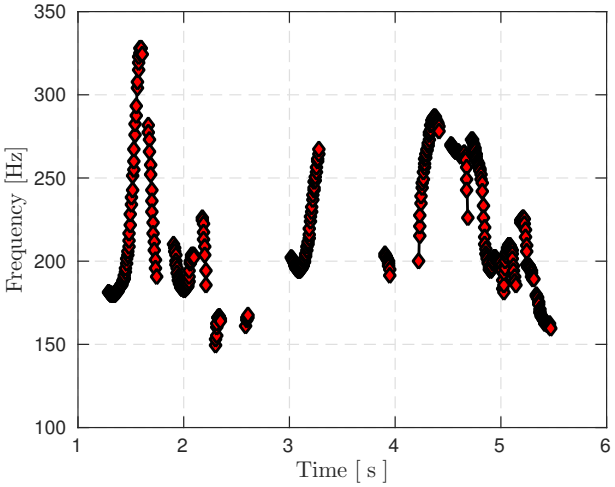


Figure 6: The estimated f_0-trajectory of speaker two uttering the (German-language) sentence [jeː neːa dɛɐ tsaɪgɐ aʊf axt kaːm dɛstɔ ʊnruːɪgɐ vʊɐdən diː lɔʏtə] (IPA).

introduced a small systematic error. Thus, we decided to make use of some prior knowledge: all speakers were walking on marked trajectories. Moreover, the visual evaluations of the videos revealed that all speakers were walking on the trajectories without noticeable deviations. Therefore, we computed the squared error between each detection and a fine grid of points on the trajectories. Then we determined the point on the trajectories where the squared error of a detection exhibited the global minimum, and mapped the detection to this point of the trajectory (see Fig. 5). We provide the original and modified skeleton tracks as text files.

2.7.3. Estimating Fundamental Frequency

First, we upsampled the glottogram to a sampling frequency of 96 kHz. Then, we filtered the signal with a Kaiser window order-estimated bandpass filter with a lower and upper cut-off frequency of 70 Hz and 8000 Hz. After compensating the introduced group-delay, we split the whole signal into frames considering a frame length of 32 ms and a frame shift of 5 ms. We computed the one-sided unbiased auto-correlation of each frame and applied a maximum detector based on the Lemire-algorithm (Lemire, 2006) with a window-size of 10 samples to each frame. After eliminating maxima with a lag below 2 ms and above 13 ms, we selected the global maximum of all remaining maxima. To eliminate outliers (e.g., caused by the speaker's act of swallowing), we computed the first derivative of the f_0-trajectory and eliminated sudden jumps with a minor sixth upwards and downwards. We provide the glottograms as wav-files and the trajectories of the estimated f_0s (see Fig. 6) as text files.

2.8. Orthographic Transcription

To generate accurate transcriptions of the recorded utterances, we followed the transcription guidelines mentioned in (Schuppler et al., 2014b), which lists all symbols used for the orthographic transcription.

3. Quality Assurance & Validation

We prepared protocols for each speaker's (sub-)session beforehand; these protocols defined what to say and where to go. They contained all selected sentences, positions, as well as orientations. During each (sub-)session, two assistants supervised the speaker by verifying the read items, the positions and orientations, and the speaker's gait velocity. We used the protocols to create a first transcription of the utterances. Afterwards, three experts checked all recorded utterances, video tracks, and text files, and made corrections if required.

4. Results on AMISCO

We evaluated our joint DOA and f_0 detector in terms of recalls and root-mean-square errors by using a subset of the AMISCO's recordings. From a set of 24 speakers, we randomly selected one male speaker and one female speaker. The evaluation's results are listed in (Pessentheiner et al., 2015) and shown as cumulative distribution functions of recalls and root-mean-square errors. (Ziegerhofer, 2016) analyzed our laryngograph (electroglottograph) recordings focusing on gender differences and speaker identity for excitation signal synthesis—the synthesis of a vocal fold's movements. (Pichler, 2016) used parts of our corpus to evaluate the performance of differential microphone arrays for speaker localization and speaker separation.

5. Discussion

During the recording and the post-processing, we encountered three problems. First, not being able to connect the camera and the Kinects to the word clock used for the audio recordings, we noticed varying delays between the starting-point of the audio and video recordings causing asynchronous video data. To overcome this problem, a person clapped his hands once in the middle of the room at the beginning and the end of each sub-session. Captured by the audio and video devices, we were able to synchronize the audio recordings with the video data during post-

processing by acoustically and visually aligning the moment of clapping in the audio and video tracks. Doing so for each recording, we realized that there was no significant drift between those two devices. Second, we had to split the 24 MEMS microphones into three groups due to the fixed number of eight microphone-connections on the microcontroller-boards. We knew that there will be clock-drifts and synchronization problems between the boards, because they were not connected to a central word clock (this was a hardware-restriction). Thus, we set up the 24-element circular array in a way that eight MEMS microphones connected to one board represent an 8-element circular array with constant angular interval of $45°$; considering all three circular arrays, the merged 24-element circular array exhibits an interval of $15°$. Third, due to an undetectable and unpredictable problem with the internal power supply of the laryngograph, speaker one exhibits distorted glottograms that should not be used. Speaker 24 doesn't include any skeleton tracks due to undetected communication problems between the Kinects and the computer during recording.

6. Availability

The website of the corpus (SPSC.tugraz.at, 2015) provides audio samples along with further information on the corpus. It will inform you about how to obtain a copy of the corpus and scripts to extract the f_0 of the glottogram, and how to process the raw skeleton tracks (in case you want to apply different algorithms to this data). Our website also provides the symbols for the orthographic transcriptions. The whole corpus will be available to research communities and institutions from summer, 2016.

7. Conclusion

The Austrian German multi-sensor corpus (AMISCO) is a collection of two-room and 43-channel close- and distant-talking Austrian German high-quality speech recordings from 24 moving and non-moving speakers, balanced male and female. It contains around 8.2 hours of read speech, $53,000$ word tokens based on $2,070$ unique word types. Furthermore, this corpus features orthographic transcriptions, glottograms, fundamental frequencies, and positions and orientations of speakers located at certain positions or walking along pre-defined trajectories. The synergy of all these different components yields a unique and comprehensive corpus that can be used in several fields of research, e.g., linguistic studies, signal processing, or machine learning. Additionally, we showed how to synchronize audio recordings with video data and how to set up Kinects that were not perfectly calibrated.

8. Acknowledgements

We acknowledge the following support: The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFW, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria, and The Technology Agency of the City of Vienna (ZIT). The programme COMET is conducted by Austrian Research Promotion Agency (FFG). The DIRHA-Project was funded by the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement No. FP7-ICT-2011-7-288121. Furthermore, this project has received funding from the Marshall Plan Foundation. The Tesla K40 GPU-cards used for the related studies were donated by the NVIDIA Corporation.

9. Bibliographical References

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, reprint (09/29/1994) edition, May.

de Cheveigné, A. and Kawahara, H. (2002). YIN, a Fundamental Frequency Estimator for Speech and Music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, Apr.

de Cheveigné, A. and Slama, M. (2006). Acoustic Scene Analysis Based on Power Decomposition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 49–52, Toulouse, May. IEEE.

DIRHA.fbk.eu. (2015). DIRHA: Distant-speech Interaction for Robust Home Applications. (accessed 02/02/2016). [Online]. Available: http://dirha.fbk.eu. Grant Agreement No. FP7-ICT-2011-7-288121.

Habib, T. and Romsdorfer, H. (2013). Auditory Inspired Methods for Localization of Multiple Concurrent Speakers. *Computer Speech and Language*, 27(3):634–659, May.

Hagmüller, M., Cristoforetti, L., and Omologo, M. (2015). DIRHA-Deliverable 2.5-2.6: Design, Collection and Transcription of Real Acoustic Corpora and Text Data (DIRHA Corpora II), Apr. Grant Agreement No. FP7-ICT-2011-7-288121.

Hänsler, E. and Schmidt, G. (2004). *Acoustic Echo and Noise Control: A Practical Approach*. Wiley, New Jersey, June.

Imoto, K., Ohishi, Y., Uematsu, H., and Ohmuro, H. (2013). Acoustic scene analysis based on latent acoustic topic and event allocation. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Southampton, September. IEEE.

Jensen, J. R., Christensen, M. G., and Jensen, S. H. (2010). Joint DOA and Fundamental Frequency Estimation Methods Based on 2-D Filtering. In *Proceedings of the 18th European Signal Processing Conference (EUSIPCO)*, pages 2091–2095, Aalborg, Denmark, August.

Jensen, J. R., Christensen, M. G., and Jensen, S. H. (2013). Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):923–933, May.

Jensen, J. R., Christensen, M. G., Benesty, J., and Jensen, S. H. (2015). Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(1):174–185, January.

Karimian-Azari, S., Jensen, J. R., and Christensen, M. G. (2013). Fast Joint DOA and Pitch Estimation Using

a Broadband MVDR Beamformer. In *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, pages 1–5, Marrakech, Morocco, September.

Kronvall, T., Adalbjörnsson, S. I., and Jakobsson, A. (2014). Joint DOA and Multi-Pitch Estimation Using Block Sparsity. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICCASP)*, pages 3958–3962, Florence, Italy, May.

Kwon, H., Krishnamoorthi, H., Berisha, V., and Spanias, A. (2009). A sensor network for real-time acoustic scene analysis. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 169–172, Taipei, May. IEEE.

Le Roux, J., Vincent, E., Hershey, J. R., and Ellis, D. P. W. (2015). Micbots: Collecting large realistic datasets for speech and audio research using mobile robots. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5635–5639, South Brisbane, Australia, Apr.

Lemire, D. (2006). Streaming Maximum-Minimum Filter Using no more than Three Comparisons per Element. *Nordic Journal of Computing*, 13(4):328–339.

Matassoni, M., Astudillo, R. F., Katsamanis, A., and Ravanelli, M. (2014). The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1613–1617, Singapore, September.

Moosmüller, S., Schmid, C., and Brandstätter, J. (2015). Standard Austrian German. *Journal of the International Phonetic Association*, 45(3):339–348, Dec.

Ngan, L. Y., Wu, Y., So, H. C., Ching, P. C., and Lee, S. W. (2003). Joint Time Delay and Pitch Estimation for Speaker Localization. In *Proceedings of the International Symposium on Circuits and Systems (ISCAS)*, volume 3, pages III–722–III–725, Bangkok, Thailand, May.

Pessentheiner, H., Hagmüller, M., and Kubin, G. (2015). Localization and Characterization of Multiple Harmonic Sources. *IEEE Transactions on Audio, Speech, and Language Processing (submitted 09/24/2015)*.

Pichler, T. C. (2016). Speaker Localization and Separation with Differential Microphone Arrays (working title). Master's thesis, Graz University of Technology, Signal Processing and Speech Communication Laboratory, Inffeldgasse 16c/EG, 8010 Graz, Styria, Austria.

Pirker, G., Wohlmayr, M., Petrik, S., and Pernkopf, F. (2011). A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1509–1512, Florence, Italy, August.

Plante, F., Meyer, G. F., and Ainsworth, W. A. (1995). A Pitch Extraction Reference Database. In *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 837–840, Madrid, Spain, Sep.

Schuppler, B., Adda-Decker, M., and Morales-Cordovilla, J. A. (2014a). Pronunciation variation in read and conversational Austrian German. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1453–1457, Singapore, Sep.

Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., and Pessentheiner, H. (2014b). GRASS: The Graz Corpus of Read and Spontaneous Speech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1465–1470, Reykjavik, Iceland, May.

SPSC.tugraz.at. (2015). Signal Processing and Speech Communication Laboratory. (accessed 10/13/2015). [Online]. Available: https://www.spsc.tugraz.at/tools/amisco.

Talkin, D., (1995). *A Robust Algorithm for Pitch Tracking (RAPT)*, pages 495–518. Elsevier Science, Amsterdam, Dec.

Teutsch, H. (2007). *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Springer, Berlin, Heidelberg, January.

Tsiami, A., Rodomagoulakis, I., Giannoulis, P., Katsamanis, A., Potamianos, G., and Maragos, P. (2014). ATHENA: A Greek Multi-Sensory Database for Home Automation Control. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1608–1612, Singapore, September.

Vary, P. and Martin, R. (2006). *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, Chichester, Jan.

DeLiang Wang et al., editors. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, New Jersey, September.

Woelfel, M. and McDonough, J. (2009). *Distant Speech Recognition*. Wiley, Chichester, Apr.

Wohlmayr, M., Stark, M., and Pernkopf, F. (2011). A Probabilistic Interaction Model for Multipitch Tracking with Factorial Hidden Markov Models. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):799–810, May.

Wrench, A. (2000). A Multi-Channel/Multi-Speaker Articulatory Database for Continuous Speech Recognition Research. In *Proceedings of the Workshop on Phonetics and Phonology in ASR (PHONUS 5)*, pages 1–14, Saarbrücken, Germany, Mar. Institute of Phonetics, Saarland University.

Ziegerhofer, J. (2016). Excitation Signal Analysis: Gender Aspects. Master's thesis, Graz University of Technology, Signal Processing and Speech Communication Laboratory, Inffeldgasse 16c/EG, 8010 Graz, Styria, Austria.