Automatic Phonetic Transcription in Two Steps: Forced Alignment and Burst Detection

Barbara Schuppler, Sebastian Grill, André Menrath, and Juan A. Morales-Cordovilla

Signal Processing and Speech Communication Laboratory, Graz University of Technology, b.schuppler@tugraz.at

Abstract. In the last decade, there was a growing interest in conversational speech in the fields of human and automatic speech recognition. Whereas for the varieties spoken in Germany, both resources and tools are numerous, for Austrian German only recently the first corpus of read and conversational speech was collected. In the current paper, we present automatic methods to phonetically transcribe and segment (read and) conversational Austrian German. For this purpose, we developed an automatic two-step transcription procedure: In the first step, broad phonetic transcriptions are created by means of a forced alignment and a lexicon with multiple pronunciation variants per word. In the second step, plosives are annotated on the sub-phonemic level: an automatic burst detector automatically determines whether a burst exists and where it is located. Our preliminary results show that the forced alignment based approach reaches accuracies in the range of what has been reported for the inter-transcriber agreement for conversational speech. Furthermore, our burst detector outperforms previous tools with accuracies between 98% and 74% for the different conditions in read speech, and between 82% and 52% for conversational speech.

Keywords: Speech Transcription, Austrian German, Conversational Speech, Automatic Burst Detection, Forced Alignment

1 Introduction

In the last decade, there was a growing interest in spontaneous and conversational speech in the fields of human and automatic speech recognition. Therefore, large conversational speech corpora have been created for many languages (e.g., for English [17], for Japanese [13], for Dutch [5]), and for French [28]). For conversational German, large speech resources are limited to the varieties spoken in Germany (e.g., [7],[31],[12]). For the varieties of Austria, only recently the first corpus of conversational speech was recorded (i.e., Graz corpus of Read and Conversational Speech (*GRASS*) [27]). In order to make the GRASS corpus accessible for speech technology as well as linguistic and phonetic research, it needs to be segmented and transcribed phonetically. The aim of the current paper is to present a transcription tool for read and conversational German. The tool is

operating in two subsequent steps. First, a broad phonetic transcription is created by means of a forced alignment (i.e., with a HMM-based approach). Second, a non-stochastic MATLAB tool annotates whether plosives are realized with a burst and, in case of an existing burst, where it is positioned. The resulting transcriptions are exported to PRAAT TextGrid format [3].

1.1 STEP 1: Broad Phonetic Transcription

Traditionally, phonetic transcriptions are produced manually by one or more transcribers. Since this approach is time consuming, methods have been developed to create broad phonetic transcriptions with the help of an ASR system (e.g., [2], [4], [9]). The accuracy of these systems has steadily increased and the agreement between automatic and manual transcriptions for some systems already is in the range of the agreements reported for human transcribers (e.g., [26]). Furthermore, automatically created transcriptions have successfully been used for phonetic investigations concerning pronunciation variation (English: [34]; German: [1]; French: [2] and Dutch: [26]).

There are different methods for creating broad transcriptions automatically. For instance, free and constrained phone recognition have been reported to work well for read speech but not for spontaneous telephone dialogues [30]. Since we aim at using the transcription tool for the casual conversations, which are part of the GRASS corpus, we did not follow this approach.

A method which does not make use of a phone recognizer based on Hidden Markov models, has been presented by Leitner et al.[11]. Their example-based approach is non-parametric and uses methods from template-based speech recognition. This tool has been trained on isolated words read by male trained Austrian speakers. Even though this tool reaches a high accuracy on carefully pronounced speech, it does not capture the variation found in spontaneous Austrian German.

Another method for creating broad phonetic transcriptions automatically is *forced alignment* (e.g., [4],[2]). For instance, the tool MAUS (Munich Automatic Segmentation) is a forced-alignment based tool which is available for German (among other languages)[20]. It works as follows: The orthographic transcription and the speech files of an utterance are uploaded to an online-tool. Then, a canonical transcription is created for each word with the Balloon tool [19]. Then, possible pronunciation variants are created based on phonological rules. Finally, an HMM based ASR system chooses the most probable pronunciation variant for each word and places the segment boundaries. We have tested this tool for our Austrian German data of the GRASS corpus and we have observed a good accuracy of the segmentation for the read speech component. For the conversational speech, however, the MAUS tool did not cover well typical characteristics of Austrian German pronunciation. For instance, MAUS annotated the alveolar fricative, which in Austrian German is typically pronounced voiceless, as voiced. Furthermore, words which tend to be reduced in spontaneous speech were not transcribed correctly. For example, the highly frequent word *ich* 'I', which in Austria is typically pronounced as [iː], was transcribed in its canonical form

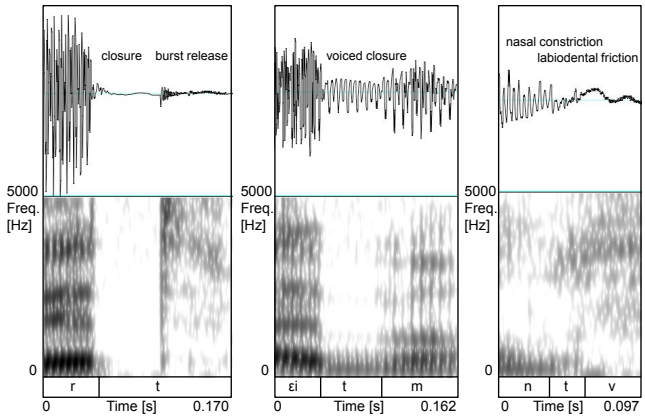

Fig. 1. Three realizations of /t/ in spontaneous Dutch. Left panel: canonical /t/. Middle panel: reduced /t/. Right panel: absent /t/[23]

/iɕ/. To conclude, none of the existing transcription tools fulfilled our requirements. Therefore, we decided to develop a HMM based ASR system in forced alignment mode to transcribe Austrian German. The main difference of our approach to the MAUS tool (described above) is that our method to creates an Austrian German pronunciation dictionary with several variants per word type (see Section 3.2). Most importantly, MAUS does not provide a sub-phonemic annotation of plosives, which is the task of Step 2.

1.2 STEP 2: Sub-phonemic Annotation of Plosives

Figure 1 shows three examples for different realizations of /t/ in conversational Dutch. The example of the left panel is the canonical realization of /t/, which consists of a voiceless complete closure followed by a strong burst and a subsequent release friction. The example on the right panel shows a realization of /t/ where all characteristic properties of a plosive are absent. [25] showed that 80.4% of /t/s in conversational Dutch are realized somewhere in between these two extremes (e.g., example in the middle panel). Recently, numerous studies investigated the different realizations of plosives in spontaneous speech and the conditions for their occurrence (for English (e.g., [18]), for Dutch (e.g., [22]), for French (e.g., [29]) and for German (e.g., [10],[35]). In these studies, sub-phonemic annotations of the plosives were created manually for a relatively small set of tokens. At the same time, they used high level statistical modeling techniques to estimate which are the predictors for the variation observed. In this paper, we propose a method to create such annotations automatically, which allows to enlarge the amount of data available for future phonetic investigations.

2 Speech Material

The speech material transcribed with the developed tool is the *Graz corpus of Read And Spontaneous Speech (GRASS)* [27]. For each of the 38 speakers (male and female), this corpus contains 62 phonetically balanced sentences, 20 (read and spontaneous) commands elicited with pictures and one hour of conversation (approximately 1200 utterances per speaker). All conversations were between family members or friends and the speakers were relaxed and talked completely freely about everyday topics (in the absence of an experimenter). Therefore, the style of the conversational speech is informal and casual. The speakers are gender balanced, with a similar average age per group. They were born in one of the eastern provinces of Austria and they all were living in Graz at the time of the recordings.

Since the corpus was collected with speech technology applications in mind, it fulfills the requirements for automatic processing (e.g., [26]): the recordings took place in a soundproof studio with both head-mounted and large-membrane microphones at 48kHz. The relative position of the speakers and the according directivity of the microphones was adjusted to optimize the SNR in the presence of overlapping speech. On average over all conversations, the resulting SNR was 46.4 dB [27].

Since for a forced alignment, the orthographic transcription is needed as input (see Section 3), the quality and consistency of the orthographic transcriptions is especially relevant. For instance, [6] reported that mistakes on the orthographic level can not be compensated on the overlying transcription layers, the contrary is the case. The orthographic transcriptions of the *GRASS* corpus were created having also such further (semi-) automatic transcription layers in mind: Speakers were transcribed on separate tiers with speech stretches of less than six seconds. Transcriptions contain information of overlapping speech, hesitations, disfluencies and other vocal and non-vocal noises [27].

3 STEP 1: Creation of a Broad Phonetic Transcription

As motivated in Section 1, we used a forced alignment to create broad phonetic transcriptions automatically. For this purpose, we used the HTK speech recognition toolkit [33]. A forced alignment needs the following input: (1) the acoustic signal, (2) the orthographic transcriptions, (3) acoustic phone models and (4) a lexicon containing pronunciation variants for each word. With this input, the alignment system determines the most likely pronunciation variant for each word appearing in the transcription of an utterance and places the corresponding segment boundaries. Finally, we exported the output of HTK to the PRAAT TextGrid format [3].

3.1 Monophone Acoustic Models

The 35 (34 phones + silence) acoustic models were continuous density 3-state monophone acoustic models with 5 Gaussians per state. The models have been

trained on 5000 utterances from 50 German speakers of the BAS read speech corpus [21]. The acoustic parameterization was as follows: 16 kHz sampling frequency, frame shift and length of 10 and 32 ms, 1024 frequency bins, 26 mel channels and 13 cepstral coefficients with cepstral mean normalization. After adding delta and delta-delta features, each final MFCC vector had 39 components (see also [24]).

3.2 Pronunciation Dictionary

The only existing pronunciation dictionary is the Austrian Phonetic Database [15]. It is based on isolated words produced by a trained speaker and thus does not cover the pronunciation variation found in the conversational speech of the *GRASS* corpus. In the following, we describe how we created a pronunciation dictionary for Austrian German, with several pronunciation variants per word type.

First, for each word a canonical pronunciation (German standard) was created with the Balloon tool [19], which makes use of a set of 49 SAMPA phoneme symbols providing syllabic and morphological boundaries, as well as primary and secondary stress. This tool is also used by the MAUS transcription system [20]. Whereas in MAUS the output is not corrected manually, we corrected the resulting canonical transcriptions. Errors mainly concerned proper names, foreign words and compounds, especially regarding the syllable boundaries and primary and secondary stress marks. The correction of the syllable boundaries and stress marks is especially important since the automatic creation of pronunciation variants is based on rules which are specific for certain syllabic structures (e.g., deletion of /r/ in coda position) and certain stress patterns (e.g., schwa deletion in unstressed syllables).

Subsequently, we applied a set of 32 rules to the canonical pronunciations. These rules can be divided into three groups. The first group is formed by those rules covering co-articulation, assimilation and reduction rules which are also typical for spontaneous German spoken by speakers from Germany. These rules include those mentioned by Wesenick et al. [32] and by Schiel, F. [20]. Secondly, we applied rules formulated on the basis of literature on standard Austrian German. Several of these rules have earlier been used for a text-to-speech engine for Austrian German [16]. The majority of these rules, however, have been formulated on the basis of phonetic studies and have not yet been used in speech technology (e.g., [14]). These rules include the deletion and lenition of plosives in all word positions. For a detailed description of each rule and their frequencies see [24].

Finally, variants were created manually for the 150 most frequent words and for certain verbs which tend to have typical Austrian realizations which cannot be easily derived from the citation form (e.g., *möchte* 'would like to': citation form /mˈœxtə/ as /mˈɛxatn/).

	Read	**Commands**	**Conversational**
Total # Phones	1826	429	10836
Deletions	0.4 % [8]	0.0 % [0]	1.3 % [133]
Insertions	1.7 % [31]	0.2 % [1]	2.1 % [228]
Substitutions	16.8 % [307]	17.0 % [73]	15.1 % [1637]
Total discrepancy	18.9 % [346]	17.2 % [74]	18.4 % [1998]

Table 1. Discrepancy between automatically created and manually corrected broad phonetic transcriptions in absolute number of phones and in % of deletions, insertions and substitutions

3.3 Validation

In order to validate the created broad phonetic transcriptions, a phonetically trained transcriber corrected the labels of the created transcriptions of part of the *GRASS* material. Then the number of substitutions, insertions and deletions was calculated (for all phones but the silence segments). Table 1 shows the discrepancies between the automatically created and the manually corrected transcription, separately for the three components of the corpus. Overall, there was a 18.5% discrepancy between the phone labels of the forced alignment and the manually corrected ones. This was mainly due to substitutions, with only a small number of insertions and deletions. These deviations between automatic and manual transcriptions are in the range of earlier reported inter-annotator discrepancies on manual transcriptions (21.2% for spontaneous speech [9]). Furthermore, the accuracy of our system lies within the range of other automatic transcription systems. For instance, [4] reported a discrepancy of 24.3% for spontaneous speech.

4 STEP 2: Automatic Sub-phonemic Annotation of Plosives

The following section describes the components of a burst-detector which is used to annotate plosives at the sub-phonemic level. The detector determines whether the plosive contains a closure and a burst and in case of a burst, it determines its position. Similar as in [8], the detector uses the power and its derivative with respect to time as principal source of information. We, however, developed a more elaborate decision stage.

4.1 Preprocessing

In a first step, the signal is Fourier transformed, high pass filtered and subsequently the power densities for each sample are accumulated to a power curve. Then, the signal passes an envelope generator that interpolates all local maxima. To suppress erroneous behavior, the interpolation stage discards all envelope points previous to the first or after the last detected maximum.

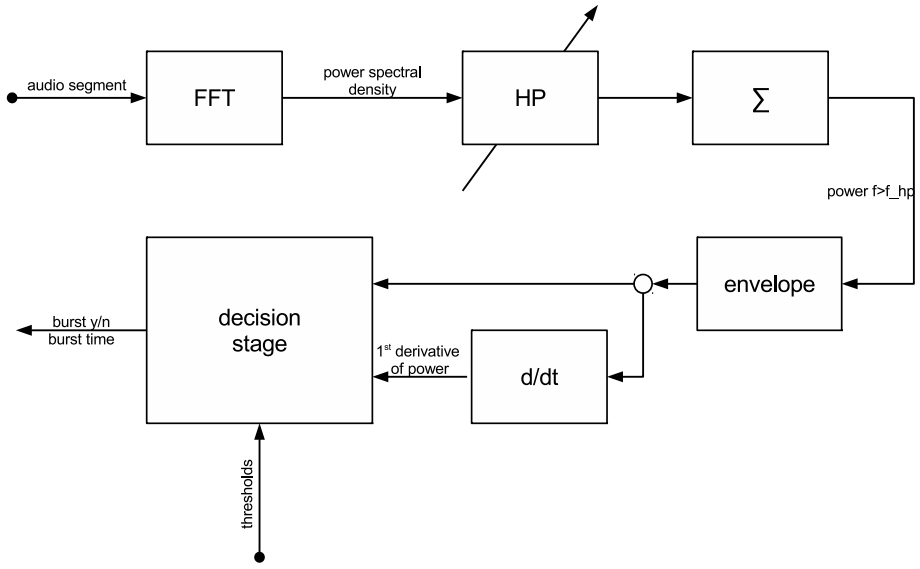

Fig. 2. *Block diagram of the complete algorithm: FFT: Fouriertransform/Spectral analysis; HP: High Pass; $\sum$: Generation of power curve from power spectral densities; d/dt: Discrete Derivative*

If insufficient supporting points are found to generate an envelope, the detection is categorically aborted and the result is set to contain no burst, as the majority of signals which result in such a condition do not contain any significant spikes in power that would hint at a burst event. We tested the positive impact of this feature on detection performance (see Section 4.3). Finally, the discrete derivative of the resulting envelope is calculated. The resulting two signals are then passed to the decision stage.

4.2 Decision Stage

First, the maximum of the derivative is compared to a plosive dependent threshold (i.e., different for /p/, /t/, /k, etc.). If the threshold is not exceeded, the decision process is aborted with the decision that no burst is present.

Second, thresholds for power as well as for the derivative are obtained by taking the maximum value of each signal multiplied by a manual parameter set. These parameters are chosen individually for each plosive. The algorithm then starts at the maximum power and proceeds backwards along the time axis until both power and its derivative fall below their respective threshold. If the values do not fall below the threshold, the decision is that no burst is present. If these two conditions are met, the process is aborted and a burst is detected.

Finally, if a burst was detected, a plosive dependent offset is added to the sample at which the algorithm stopped to obtain the burst time. The reason for

this offset stems from the usage of an envelope in the power signal, which shifts the onset of the burst forward, as well as there being an optimization problem between overall burst detection and temporal precision. If parameters are optimized to obtain optimal burst detection, the temporal precision suffers and vice versa. We found that an offset was an easy method to avoid this optimization conflict.

4.3 Sub-phonemic Annotation of Plosives in Read Speech

For evaluating the accuracy of the burst detector in read speech, we used a subset of the German *Kiel Corpus* [7]. The subset contains German read speech of the same text spoken by nine male and seven female speakers. The corpus comes with detailed manually created phonetic transcriptions, also at the sub-phonemic level of plosives, which made it an ideal reference to validate the burst detector. In total, we used 1579 bursts for the validation of our automatically created sub-phonemic annotation of the plosives.

We evaluated both the decision of the detector (is there a burst?, yes or no) and the position of the plosive (distance from manual burst in ms). For the analysis, we calculated the deviations from the manual transcription separately for the different plosives (/p/,/t/,/k/,/b/,/d/,/g/) and we grouped them in terms of position within the word (word initial, word medial and word final). For each of these combinations the following measures were calculated to estimate the accuracy of the burst detector:

- $\mathbf{P}_1$**:** A burst was detected and it was present in the manual transcription.
- $\mathbf{P}_2$**:** No burst was detected and it was absent in the manual transcription.
- **P:** Detector decision is correct.
- $\boldsymbol{\Delta}$**b**: Arithmetic mean of the temporal error between the detected and the manually labeled burst position (in numbers of samples).

Table 2 show the results separately for the different plosives as well as the overall result. The numbers in the squared brackets represent the number of occurrences of the respective case. In 161 cases spanning all plosives in all possible positions, no burst was detected because insufficient supporting points were found to generate a hull curve (also see section 4.1). This detector decision was correct in 98% of the cases.

Overall plosive categories, the decision of whether a burst was present was correct in 93% of the cases, with a maximum of 98% for initial, voiceless plosives and a minimum of 74% for word-final /k/. These values are much better than previously reported. [8], for instance, reached a similarly high maximum of 97% agreement for the presence of bursts in word-initial position, but only 47% agreement for the absence of bursts in word-medial position.

4.4 Sub-phonemic Annotation of Plosives in Austrian German

In order to evaluate the accuracy of the burst detector on Austrian German, we extracted 2071 word tokens containing a plosive from the *GRASS* corpus. For

	Total	/p/	/t/	/k/	/b/	/d/	/g/
Overall							
P	0.93 [1579]	0.96 [50]	0.96 [673]	0.92 [119]	0.92 [112]	0.89 [494]	0.87 [131]
$\mathbf{P}_1$	0.91 [822]	0.92 [12]	0.92 [263]	0.91 [93]	0.97 [35]	0.89 [351]	0.88 [68]
$\mathbf{P}_2$	0.95 [757]	0.97 [38]	0.98 [410]	0.92 [26]	0.90 [77]	0.91 [143]	0.86 [63]
$\Delta\mathbf{b}$	43 [744]	76 [11]	50 [243]	44 [85]	50 [34]	37 [311]	35 [60]
Initial							
P	0.92 [528]	-	0.98 [48]	1.00 [16]	0.94 [64]	0.92 [352]	0.88 [48]
$\mathbf{P}_1$	0.93 [407]	-	-	1.00 [16]	1.00 [33]	0.93 [313]	0.89 [45]
$\mathbf{P}_2$	0.90 [121]	-	0.98 [48]	-	0.87 [31]	0.85 [39]	0.67 [3]
$\Delta\mathbf{b}$	39 [379]	-	-	25 [16]	51 [33]	38 [290]	37 [40]
Medial							
P	0.92 [801]	0.97 [32]	0.96 [424]	0.95 [80]	0.90 [48]	0.82 [134]	0.87 [83]
$\mathbf{P}_1$	0.89 [338]	1.00 [6]	0.94 [208]	0.95 [61]	0.50 [2]	0.55 [38]	0.87 [23]
$\mathbf{P}_2$	0.94 [463]	0.96 [26]	0.98 [216]	0.95 [19]	0.91 [46]	0.93 [96]	0.87 [60]
$\Delta\mathbf{b}$	46 [302]	26 [6]	48 [196]	50 [58]	38 [1]	27 [21]	32 [20]
Final							
P	0.94 [250]	0.94 [18]	0.96 [201]	0.74 [23]	-	1.00 [8]	-
$\mathbf{P}_1$	0.82 [77]	0.83 [6]	0.85 [55]	0.69 [16]	-	-	-
$\mathbf{P}_2$	0.99 [173]	1.00 [12]	0.99 [146]	0.86 [7]	-	1.00 [8]	-
$\Delta\mathbf{b}$	58 [63]	135 [5]	54 [47]	39 [11]	-	-	-

Table 2. Automatic annotation of bursts in plosives in the Kiel Corpus of Read Speech. Percentages P_1 - P_3: P_1: Burst detected and it was present in the manual transcription. P_2: No burst detected and it was absent in the manual transcription. P: Detector decision is correct. $\Delta\mathbf{b}$: Temporal error (in numbers of samples, the sampling frequency was 44 kHz)

these tokens, the bursts in the plosives were annotated manually by a trained transcriber. The results for the different plosives are shown in Table 3. Since basically all tools work better for read than for spontaneous speech, it could be expected that also our burst detector did not achieve as high accuracies for the material from the *GRASS* corpus as for the carefully pronounced speech from the Kiel Corpus. Nevertheless, the tool reached a maximum accuracy of 82% for /g/ and a minimum for /b/ of 52%. These values are still within the range of what [8] (max. 97%, min. 47%), and [23] (average 63%) reported for spontaneous American English.

One explanation for the lower accuracy reached for detection of bursts in /b/ might be that /b/ is frequently realized as voiced labiodental fricative in spontaneous Austrian German. Another reason might be the different recording conditions of the two corpora. In future work, we will develop automatic methods to optimize the parameterization for the different plosives, specifically for conversational speech.

	/p/	/t/	/k/	/b/	/d/	/g/
P	0.59 [144]	0.68 [917]	0.81 [198]	0.52 [158]	0.67 [466]	0.82 [188]
$\mathbf{P}_1$	0.39 [95]	0.60 [676]	0.84 [176]	0.26 [102]	0.63 [355]	0.82 [137]
$\mathbf{P}_2$	0.98 [49]	0.90 [241]	0.64 [22]	0.98 [56]	0.82 [111]	0.82 [51]
$\Delta\mathbf{b}$	4 [37]	13 [403]	9 [147]	5 [27]	10 [222]	9 [112]

Table 3. Automatic annotation of bursts in plosives in the GRASS corpus. Percentages P_1 - P_3: P_1: Burst detected and it was present in the manual transcription. P_2: No burst detected and it was absent in the manual transcription. P: Detector decision is correct. $\Delta\mathbf{b}$: Temporal error (in numbers of samples; the sampling frequency was 16kHz)

5 Conclusions

In the current paper, we presented automatic methods to phonetically transcribe and segment the recently collected *GRASS* corpus, which is the first corpus of read and conversational Austrian German [27]. For this purpose, we developed a two-step procedure: In the first step, broad phonetic transcriptions were created by means of a forced alignment and a lexicon with multiple pronunciation variants per word. In order to create pronunciation variants typical for Austrian German, we applied 32 rules to the canonical pronunciations of the words. In a second step, all plosives were annotated on the sub-phonemic level: a burst detector automatically determined whether a burst existed in a plosive and where it was located. After this step, both the broad phonetic transcription and the sub-phonemic plosive annotation are exported in form of a PRAAT TextGrid.

The quality of both steps was evaluated separately by comparison with manually created transcriptions. We found that the forced alignment based approach reached accuracies in the range of what has been reported for inter-transcriber agreement for conversational speech. Furthermore, our burst detector outperformed previous tools with accuracies between 98% and 74% for the different conditions in read speech, and between 82% and 52% for conversational speech.

In future work, we will tune the parameters of the burst detector for the conversational speech. Then, we will use the created annotations to model which are the predictors for plosive reduction in read and conversational Austrian German in comparison to the varieties spoken in Germany. These plosive-reduction models will not only inform linguists interested in conversational speech, but they will also be incorporated into the pronunciation model of an ASR system for Austrian German.

Acknowledgements

The work by Barbara Schuppler was supported by a Hertha-Firnberg grant from the FWF (Austrian Science Fund). The work of Juan A. Morales-Cordovilla was funded by the European project DIRHA (FP7-ICT-2011-7-288121).

References

1. Adda-Decker, M., Lamel, L.: Modeling reduced pronunciations in German. Phonus 5, Institute of Phonetics, University of the Saarland pp. 129–143 (2000)
2. Adda-Decker, M., Snoeren, N.D.: Quantifying temporal speech reduction in French using forced speech alignment. Journal of Phonetics, in press (2011)
3. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 314–345 (2001), http://www.praat.org, last viewed 25-3-2014
4. Cucchiarini, C., Binnenpoorte, D.: Validation and improvement of automatic phonetic transcriptions. In: Proceedings of ISCLP. pp. 313–316. Denver, USA (2002)
5. Ernestus, M.: Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface. Ph.D. thesis, LOT, Vrije Universiteit Amsterdam, The Netherlands (2000)
6. Gubian, M., Schuppler, B., van Doremalen, J., Sanders, E., Boves, L.: Novelty detection as a tool for automatic detection of orthographic transcription errors. In: Proceedings of the 13-th International Conference on Speech and Computer SPECOM-2009. pp. 509–514 (2009)
7. IPDS: CD-ROM: The Kiel Corpus of Spontaneous Speech, vol i- vol iii. Corpus description available at http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html (last viewed 25/11/2012) (1997)
8. Khasanova, A., Cole, J., Hasegawa-Johnson, M.: Assessing reliability of automatic burst location. In: Proceedings of Interspeech (2009)
9. Kipp, A., Wesenick, M., Schiel, F.: Pronunciation modeling applied to automatic segmentation of spontaneous speech. In: Proceedings of Eurospeech. pp. 1023–1026 (1997)
10. Kuzla, C., Ernestus, M.: Prosodic conditioning of phonetic detail in German plosives. Journal of Phonetics 39, 143–155 (2011)
11. Leitner, C., Schickbichler, M., Petrik, S.: Example-based automatic phonetic transcription. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). pp. 3278–3284 (2010)
12. Lücking, A., Bergman, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Proceedings of LREC 2010 Workshop: Multimodal Corpora- Advances in Capturing, Coding and Analyzing Multimodality. pp. 92–98 (2010)
13. Makimoto, S., Kashioka, H., Nick, C.: Tagging structure and relationships in a Japanese natural dialogue corpus,. In: Proceedings of Interspeech. pp. 912–917 (2007)
14. Moosmüller, S.: The process of monophthongization in Austria (reading material and spontaneous speech). In: Papers and Studies in Contrastive Linguistics. pp. 9–25 (1998)
15. Muhr, R.: Österreichisches Aussprachewörterbuch – Österreichische Aussprachedatenbank. Peter Lang Verlag, Frankfurt/M., Wien u.a. 525 S. mit DVD (2007)
16. Neubarth, F., Pucher, M., Kranzler, C.: Modeling Austrian dialect varieties for TTS. In: Proceedings of Interspeech. pp. 1877–1880 (2008)
17. Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W.D.: The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. Speech Communication 45, 89–95 (2005)
18. Raymond, W.D., Dautricourt, R., Hume, E.: Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical and phonological factors. Language Variation and Change 18, 55–97 (2006)

19. Reichel, U.D.: PermA and Balloon: Tools for string alignment and text processing. In: Proceedings of Interspeech 2012. p. paper no. 346 (2012)
20. Schiel, F.: Automatic phonetic transcription of non-prompted speech. In: Proceedings of the ICPhS 1999. pp. 607–610 (1999)
21. Schiel, F., Baumann, A.: Phondat1, corpus version 3.4. Intern. report, http://www.bas.uni-muenchen.de/bas/basformatseng.html, Bavarian Archive for Speech Signals (BAS) (2006)
22. Schuppler, B., van Dommelen, W., Koreman, J., Ernestus, M.: How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. Journal of Phonetics 40, 595–607 (2012)
23. Schuppler, B.: Automatic Analysis of Acoustic Reduction in Spontaneous Speech. Ph.D. thesis, Radboud University Nijmegen, The Netherlands (2011)
24. Schuppler, B., Adda-Decker, M., Morales-Cordovilla, J.A.: Pronunciation variation in read and conversational Austrian German. In: Accepted for publication at Interspeech'14 (2014)
25. Schuppler, B., van Dommelen, W., Koreman, J., Ernestus, M.: Word-final [t]-deletion: An analysis on the segmental and sub-segmental level. In: Proceedings of Interspeech. pp. 2275–2278 (2009)
26. Schuppler, B., Ernestus, M., Scharenborg, O., Boves, L.: Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. Journal of Phonetics 39, 96–109 (2011)
27. Schuppler, B., Hagmüller, M., Morales-Cordovilla, J.A., Pessentheiner, H.: GRASS: The Graz corpus of Read and Spontaneous Speech. In: Proceedings of LREC'14. pp. 1465–1470 (2014)
28. Torreira, F., Adda-Decker, M., Ernestus, M.: The Nijmegen Corpus of Casual French. Speech Communication 52(3), 201–212 (2010)
29. Torreira, F., Ernestus, M.: Probabilistic effects on French [t] duration. In: Proceedings of Interspeech. pp. 448–451 (2009)
30. Van Bael, C.: Validation, Automatic Generation and Use of Broad Phonetic Transcriptions. Ph.D. thesis, Radboud Universiteit Nijmegen, Nijmegen (October 2007)
31. Weilhammer, K., Reichel, U., Schiel, F.: Multi-tier annotations in the Verbmobil Corpus. In: Proceedings of LREC. pp. 912–917 (2002)
32. Wesenick, M.B.: Automatic generation of German pronunciation variants. In: Proceedings of the ICSLP. pp. 125–128 (1996)
33. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (v. 3.2). Tech. rep., Cambridge University. Engineering Department (2002)
34. Yuan, J., Liberman, M.: Investigating /l/ variation in English through forced alignment. In: Proceedings of Interspeech. pp. 2215–2218 (2009)
35. Zimmerer, F., Scharinger, M., Reetz, H.: When BEAT becomes HOUSE: Factors of word final /t/-deletion in German. Journal of Phonetics 39, 143–155 (2011)