Timo Baumann (Hrsg.)
Elektronische Sprachsignalverarbeitung 2024
Tagungsband der 35. Konferenz
Regensburg, 6.-8. März 2024

TUDpress

Studientexte zur Sprachkommunikation
Hg. von Rüdiger Hoffmann
ISSN 0940-6832
Bd. 107

Timo Baumann (Hrsg.)

Elektronische Sprachsignalverarbeitung 2024

Tagungsband der 35. Konferenz

Regensburg, 6.-8. März 2024

TUDpress

2024

Einbandfoto, Hintergrund: Mitch Rue
Einbandfoto Zentralmotiv: Verena Willkomm

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

ISBN 978-3-95908-325-6

D-01309 Dresden
Tel.: +49 351 4721463
http://www.tudpress.de

TUDpress ist ein Imprint von Thelem

Gesetzt von den Herausgebern.
Printed in Germany.

Vorwort

Liebe Freunde der Tagungsreihe, liebe Kolleginnen und Kollegen,

die 35. Konferenz „Elektronische Sprachsignalverarbeitung" (ESSV) findet in diesem Jahr erstmals an der Ostbayerischen Technischen Hochschule Regensburg statt. Ich freue mich auf die vielfältigen Beiträge, die Gelegenheiten zum persönlichen Gespräch und die daraus resultierenden wissenschaftlichen Austausche. Und auch auf die Gelegenheit alte Bekanntschaften zu pflegen sowie neue aufzubauen.

Die Ausrichtung der ESSV ist für mich eine besondere Ehre. Die Tatsache, dass sie ein zweites Mal hintereinander im Freistaat Bayern – aber erstmals überhaupt in der Oberpfalz! – stattfindet, ist meinem unsteten akademischen Lebenswandel geschuldet. Mein Dank gilt dem Förderverein Elektronische Sprachsignalverarbeitung e.V. für das Vertrauen, die diesjährige ESSV in meine Hände zu legen und insbesondere auch dafür, mich bei der Organisation beständig so freundlich und liebevoll drängelnd zu unterstützen.

Ich freue mich besonders, dass in diesem Jahr die ESSV an einer Technischen Hochschule anstatt wie üblich an einer Universität stattfindet. Die praxisnahe Ausrichtung der ESSV und die Verzahnung mit Industrie und außeruniversitären Forschungseinrichtungen trifft hier auf die weiter wachsende Forschungsstärke der Fachhochschulen die sich für die OTH Regensburg seit diesem Jahr im eigenen Promotionsrecht zeigt.

Auch zu dieser ESSV hat es wieder viele Einreichungen gegeben, die in insgesamt 29 begutachteten und angenommenen Beiträgen resultierten, die das ganze Themenspektrum der elektronischen Sprachsignalverarbeitung abdecken und in diesem Tagungsband versammelt sind. Hinzu kommen zwei spannende Hauptvorträge: Anna Kruspe thematisiert mit der Spracherkennung für Gesang einen besonders kunstvollen Teil sprachlicher Äußerungen und Hendrik Buschmeier spricht über Höflichkeit als besonders wertvollen Aspekt des sprachlichen Miteinanders mit Menschen und Maschinen. Gemeinsam mit der ESSV findet zum zweiten Mal ein Workshop der Informationstechnischen Gesellschaft zu Sprachassistenten statt, welcher weitere Beiträge zu diesem wichtigen Thema umfasst.

Eine Neuerung über die ich mich freue ist die dauerhafte digitale Bereitstellung der diesjährigen Beiträge mittels DOI und URN. Hierfür danke ich der Hochschulbibliothek der OTH. Auch in vielem anderen bin ich der OTH Regensburg zum Dank verpflichtet – so ist die Ausrichtung der ESSV ja nicht bloß eine Ehre sondern auch ein gewisser Aufwand und finanzielles Risiko. Sehr danken möchte ich in diesem Zusammenhang unseren Sponsoren, Genie Enterprises sowie Alphaspeech, ohne die die ESSV in dieser Form nicht möglich wäre.

Regensburg, im Februar 2024
Timo Baumann

Inhaltsverzeichnis

Eingeladene Vorträge

Prof. Dr.-Ing. Anna Kruspe, Hochschule München

More Than Words: Advancements and Challenges in Speech Recognition for Singing

Prof. Dr.-Ing. Hendrik Buschmeier, Universität Bielefeld

Linguistic Politeness in Artificial Conversational Agents

MORE THAN WORDS: ADVANCEMENTS AND CHALLENGES IN SPEECH RECOGNITION FOR SINGING

Anna Kruspe
University of Applied Sciences Munich
anna.kruspe@hm.edu

Abstract: This paper addresses the challenges and advancements in speech recognition for singing, a domain distinctly different from standard speech recognition. Singing encompasses unique challenges, including extensive pitch variations, diverse vocal styles, and background music interference. We explore key areas such as phoneme recognition, language identification in songs, keyword spotting, and full lyrics transcription. I will describe some of my own experiences when performing research on these tasks just as they were starting to gain traction, but will also show how recent developments in deep learning and large-scale datasets have propelled progress in this field. My goal is to illuminate the complexities of applying speech recognition to singing, evaluate current capabilities, and outline future research directions.

1 Introduction

Automatic Speech Recognition (ASR) technology has seen significant advancements in various domains but remains relatively underexplored in the field of singing. This gap in research may initially appear justifiable, given the widespread availability of song lyrics. However, the importance of ASR in singing extends far beyond mere convenience. It is crucial in niche music genres and less mainstream tracks, often described as 'long tail' music, where lyrics are not readily available. This technology also aids in making music more accessible to the hearing impaired through real-time captioning and can be an engaging tool for language learning and cultural exchange.

The necessity for this technology becomes more pronounced in the context of world music, a genre characterized by its diverse linguistic and cultural roots. Here, speech recognition can play a crucial role in breaking down language barriers and promoting cultural understanding. It can also enhance karaoke and entertainment applications, improve music search and discovery, and aid in music analysis and research.

Moreover, the future of AI models in this domain promises a more efficient approach to lyrics transcription. Current methods, often relying on user submissions or manual transcription, are time-consuming and can be inaccurate. Advanced speech recognition could potentially automate this process, providing quicker and more accurate transcriptions, and assist in content moderation for public performances.

The scope of ASR in singing extends beyond transcription. It includes practical applications like synchronizing lyrics with audio, identifying songs from sung lyrics (useful in platforms like music identification apps), transcribing lyrics from short audio clips, identifying the language of the song, and enhancing music recommendation systems by understanding lyrical content and sentiments.

My work in ASR for singing began in 2011, a time when the field was just starting to embrace new solutions for its challenges. This period also marked the rise of deep learning methods in various scientific areas, significantly impacting the field of ASR. My research coincided

with this shift from traditional feature-based methods to more advanced, data-driven neural networks. The transition was particularly notable in ASR for singing. Earlier models, which were primarily designed and trained for regular speech, often underperformed with singing due to a lack of specialized adaptation techniques and data. The newer deep learning models have proven to be more effective in this regard. We are now beginning to see solutions that are not only theoretically sound but also practically viable for addressing the unique challenges of ASR in singing.

This paper aims to provide an overview of speech recognition specifically tailored to singing, including my own forays into the topic. It will begin with an examination of the unique challenges that singing poses as a data source, such as varied vocal styles, pitch variations, and background music interference. Following this, the paper will delve into specific research areas: phoneme recognition in singing (critical for understanding lyrics), sung language identification (which can be particularly challenging given the musical context), keyword spotting in songs (useful for searching and categorizing music), and methods for retrieving songs based on the sung lyrics. Additionally, we will review the progress in the complete transcription of songs, assessing current capabilities and limitations. In most cases, I will give a historical overview of first approaches, my own work between 2011 and 2018, and recent developments.

Finally, the paper will conclude with a summary of the key findings and a discussion on future research directions. This will include potential technological advancements, the integration of these systems into consumer applications, and the exploration of new use cases in the rapidly evolving landscape of music and technology.

2 Singing as a speech data source

Singing poses several unique challenges for speech recognition compared to normal speech. These challenges necessitate adapting existing speech recognition algorithms [1]:

Larger Pitch Fluctuations Singing involves more significant pitch variations than speaking, along with different spectral properties.

Increased Loudness Variability Loudness in singing fluctuates more than in speech.

Pronunciation Variation The musical context can lead singers to pronounce sounds and words differently compared to normal speech.

Time Variations In singing, sounds may be elongated or shortened to fit the musical rhythm, leading to more significant variations, especially with vowels. This was confirmed by a study comparing standard deviations of phonemes in speech and singing datasets.

Different Vocabulary Lyrics in songs often use different words and phrases compared to everyday conversation, with a focus on emotional topics.

Background Music Interference In polyphonic recordings, the presence of harmonic and percussive instruments adds spectral components that confuse speech recognition algorithms. Source separation algorithms ould be utilized to remove these components, but they are not always effective and can introduce artifacts. Vocal Activity Detection can be used to at least discard non-singing segments in songs, but it often makes errors in problematic cases for speech recognition, such as instrumental solos. Due to these challenges, most works focus on unaccompanied singing and leave the pre-processing as a separate step to be researched. Alternatively, algorithms are being developed with robustness to these influences in mind from the start.

Historically, ASR for singing faced significant challenges due to a lack of available data. Unlike speech data, which is often created specifically for ASR development, music and lyrics are generally subject to copyright restrictions that limit their use. Additionally, phoneme- or

word-level annotations, crucial for ASR training, are typically scarce, and funding for creating such annotations is limited. As a result, early ASR models for singing were primarily adapted from speech data models and tested on small, available singing datasets.

When I began my research in this field, only two modest-sized datasets were available, each containing around 20 English-language pop songs with line-wise lyric annotations [2, 3]. I then utilized a dataset from *Smule's* amateur karaoke app *Sing!*, known as *DAMP*, which comprised unaccompanied singing. Initially lacking lyric labels, I compiled lyrics from the *Smule* website and conducted forced alignment using speech-trained models, which led to the creation of a set with 300 songs, each having 20 recordings [4]. Subsequent work improved the phonetic annotations of this dataset and expanded its scope [5], making it a standard resource for ASR in singing.

The availability of data has significantly improved since then. New datasets have been introduced, such as *MUSDB* [6], containing 150 mostly English-language songs with line-wise lyrics annotations, and *vocadito* [7], featuring 40 manually labeled multilingual recordings. Another notable dataset is *DALI* [8, 9], which includes thousands of polyphonic recordings in various languages, with lyrics annotations obtained through semi-automated methods.

The recent proliferation of these datasets makes direct comparison with older approaches challenging due to the lack of established benchmarks. Therefore, in the following sections, my focus will be on exploring individual ideas for the tasks discussed, rather than providing numerical comparisons.

3 Phoneme recognition for singing

Phoneme recognition in singing, also known as acoustic modeling, has traditionally been more challenging than in speech. This is due to the unique characteristics of singing described above, such as varied pitch and rhythm. For a long time, phoneme recognition was foundational for other ASR tasks like alignment, making it a critical area of study. Recognizing phonemes in singing is also a complex task for humans, as noted in [10].

Early phoneme recognition systems, which often relied on Mel-Frequency Cepstral Coefficients (MFCCs) and assumed pitch invariance, faced difficulties in accurately processing singing. Two initial systems using Hidden Markov Models (HMMs) were introduced in [11, 12]. These were followed by approaches involving adapted Gaussian Mixture Model-HMMs [13] and systems utilizing chorus repetitions [14]. However, a lack of benchmark datasets for singing made direct comparisons with these early systems challenging.

The scarcity of large-scale, singing-specific training data also hindered progress. My own research initially involved using speech datasets like *TIMIT* [15], but resulted in high error rates. To improve model robustness, I experimented with augmenting speech data to mimic singing characteristics, such as pitch shifting and time stretching [16]. This led to reduced error rates and later informed the integration of these techniques into advanced models using Transformer architectures [17].

Subsequently, I worked with the *DAMP* dataset, derived from *Smule Sing!* karaoke app recordings. By creating phoneme labels through forced alignment with TIMIT models, I demonstrated that direct training on singing data was significantly more effective [4].

Presently, the focus in singing ASR has shifted towards full transcription, which involves integrating acoustic models with language models or employing end-to-end systems (also see section 8). Nonetheless, the insights and methodologies from phoneme recognition research continue to be valuable for understanding singing's unique phonetic characteristics.

4 Sung language identification

Several methodologies have been explored for language identification in singing. In 2004, an unsupervised clustering approach was developed to create language-specific codebooks from input features, achieving an accuracy of 0.8 for English and Mandarin songs [18]. [19] in 2006 utilized MFCC features for direct language model training, but faced challenges in singing and polyphonic contexts, indicating the complexity of the task . [20] in 2011 combined phoneme recognition with prosodic tokenization, testing on a multilingual corpus and achieving accuracies up to 0.83. In the same year, [21] analyzed audio and video features in music videos, noting that identifying European languages was more challenging than Asian and Arabic languages, with accuracies around 0.45 using audio and 0.48 with both audio and video.

My work began with systems based on traditional audio features like MFCC and RASTA-PLP, fed into machine learning models [22]. I later incorporated the i-vector technique, primarily used in speaker recognition, for feature reduction [23]. A notable observation was the models' tendency to confuse speaker characteristics with language features. Subsequently, I employed phoneme statistics for language identification, building on the phoneme recognition approaches described above [24]. The phonotactic approach was subsequently also taken in [25].

Recent developments have seen the adoption of advanced neural networks, along with the integration of auxiliary textual data such as song, artist, and album names [26]. The availability of large multilingual datasets has also improved, easing the research process [27]. Despite these advancements, language identification in singing remains a challenging task, particularly for languages with limited resources in singing data.

5 Keyword spotting

Keyword-based search systems play a crucial role in various music-related applications, such as song discovery based on topics, playlist creation, similarity searches, genre classification, and mood detection. Early methods often relied on supplementary information like textual lyrics. For instance, a 2008 study employed vocal re-synthesis with MFCCs and power features for phoneme recognition, using Viterbi decoding alongside keyword-filler HMMs [28]. In 2016, the "LyricListPlayer" system utilized lyrics-to-audio alignment for keyword detection, incorporating NLP techniques for topic modeling [29]. [30] presented an approach involving Statistical Sub-Sequence DTW for keyword spotting, which required audio recordings of key phrases. Additionally, [31] developed a score-aided method that combined acoustic keyword spotting with Sub-Sequence DTW and Dynamic Bayesian Network HMMs, tested on Turkish Makam music.

In my research, I focused on detecting arbitrary keywords in singing without needing extra information. The primary approach involved using keyword-filler HMMs, which consist of several phoneme-level states for detecting the desired keyword and an additional state for all other sounds [32]. This method was further refined by incorporating knowledge about plausible phoneme durations to eliminate unlikely candidates [33].

Recently, keyword spotting in singing has seen less research focus, possibly due to the anticipation that comprehensive transcription systems may render individual keyword searches obsolete (refer to section 8). Nevertheless, in scenarios where complete lyrics are unavailable or transcription systems are not fully accurate, keyword spotting methods retain their relevance. They offer the ability to identify more potential keyword instances within songs under model uncertainty.

6 Lyrics alignment

Lyric-to-audio alignment has been a more extensively researched topic compared to other areas discussed earlier. [34] first adapted speech recognition methods for singing using MFCCs and a modified Viterbi algorithm in a HMM for unaccompanied singing in 1999. However, this early approach was limited by its small database and lack of quantitative results. In 2004, a significant advancement was made with "LyricAlly," providing line-level alignments in polyphonic recordings, employing a mix of rhythm structure analysis, chord analysis, chorus detection, Vocal Activity Detection (VAD) through HMMs, and lyric segmentation [35].

Further progress was made in 2006, focusing on syllable-level alignment and adapting speech acoustic models for singing. This period saw methods like auto-regressive HMMs for modeling high-pitched signals and MFCC-based Viterbi alignment with accompaniment sound reduction and phoneme model adaptation [36]. In 2010, [37] introduced the use of chord labels to improve alignment accuracy [37]. Other specialized approaches included a 2007 method for Cantonese singing, utilizing prosodic information from lyrics [38], and structural analysis of song recordings in 2008 [39]. [40] in 2008 incorporated harmonic re-synthesis for vocal separation, and [41] in 2015 focused on vowel alignment in score-following algorithms. [42] in 2016 enhanced accuracy by integrating note onsets into the alignment algorithm.

In my work, I applied Dynamic Time Warping (DTW) to align phoneme posteriorgrams, derived from phoneme recognition, with binary templates generated from lyrics. This approach was further refined with insights into phoneme probabilities and confusions, enhancing alignment accuracy. In the 2017 *MIREX* challenge, this method achieved the lowest mean error rates in both unaccompanied and polyphonic music [43].

The annual *MIREX* challenge has improved the reproducibility and comparability of alignment methods. Recent developments include enhancements to acoustic modeling [44], an end-to-end solution using Wave-U-Net and CTC loss [45], extensions to multilingual data [46], and a computationally efficient approach for detecting anchor points [47]. Recently, the focus has shifted towards combining alignment with full transcription (see section 8).

7 Lyrics-based retrieval

Lyrics-based retrieval, a relatively underexplored area in ASR research, involves identifying the correct textual lyrics and corresponding songs from a sung query. This technology is particularly beneficial for karaoke systems and voice-based search applications.

In 2006, [48] developed a phoneme recognition system for lyrics retrieval. Their experiments showed that using a five-word query improved the retrieval rate, and integrating melody recognition further increased its accuracy. The query-by-singing system from [49] combined melody and lyrics in 2010, while [50] developed a system focused solely on lyrics for word recognition in singing.

In my research, I utilized the outcomes of the phoneme recognition process described above, bypassing additional language modeling or melody integration, on a database containing 300 songs [51]. Sung lines from these songs served as retrieval queries. Adaptations made to the alignment process, as detailed earlier (section 6), were instrumental in accommodating variations in recognized phonemes and common confusions [52].

Recent years have seen limited focus on lyrics-based retrieval research. Full transcription capabilities could potentially address this task by enabling a fuzzy text search within a lyrics database. Nevertheless, integrating these lyrics-based methods with audio-based song identification could significantly enhance song search capabilities, especially for cover versions or queries by amateur singers who recall only fragments of the melody and lyrics.

8 Transcription

Lyrics transcription has long been viewed as the "holy grail" in the field of ASR for singing. Traditionally, an ASR pipeline involved an acoustic model to determine phoneme likelihoods and a language model to deduce the most probable sequences of phonemes and words, informed by statistical information from text data sets. However, in the context of singing, as noted in section 3, the acoustic models initially struggled to provide sufficient accuracy for full transcription. [53] introduced deep learning-based (TDNN-LSTM) acoustic models trained on limited singing data in 2018.

With the availability of larger datasets and advancements in ASR technology, the pursuit of lyrics transcription has intensified recently. [44] explored an end-to-end model combining acoustic and language modeling, although it initially underperformed compared to separate models. [54] then employed Time Delay Neural Networks (TDNNs) integrated with language models trained on diverse lyrics data, achieving improved error rates. Their subsequent work adopted a multistream approach for further enhancements [55].

More recently, end-to-end transcription has become feasible with the adoption of Transformer and Conformer architectures. For instance, [56] extended these models with genre-specific adapters. [57] successfully adapted wav2vec embeddings for singing, marking a significant leap in model performance. In the latest development, a novel approach involved performing ASR on music audio with the *Whisper* system, followed by post-processing using *ChatGPT*, leading to further reductions in error rates [58]. It is anticipated that future research will increasingly leverage such versatile, pre-existing systems or adapt large language models (LLMs) to this domain.

9 Conclusion and future work

This paper has examined the challenges and recent progress in speech recognition for singing. We have seen how singing's unique characteristics, like varying pitches and durations, complex pronunciations, and background music, present different challenges from regular speech recognition. Significant advances have been made in key areas such as phoneme recognition, sung language identification, keyword spotting, and full song transcription. In recent years, improvements are largely due to advancements in deep learning and the availability of diverse, large datasets.

The transition from feature-based methods to deep learning signifies a major change in this field. These advanced models are better at understanding the subtleties of singing, leading to more precise and reliable ASR systems. Research exploring various languages and music styles continues to expand, enriching our understanding of music from around the world.

The potential impact of these developments on music discovery and recommendation systems is substantial. More accurate song identification and transcription can lead to better, more varied music suggestions, helping users discover new artists and genres. This not only improves the listening experience but also supports lesser-known music. Additionally, these advances in ASR can make music from different cultures more accessible and help overcome language barriers.

Looking forward, the focus is likely to shift more towards complete transcription of songs. The success in transcription could indirectly solve related tasks such as keyword spotting and lyrics alignment, offering a unified solution to several challenges in the domain. Moreover, the advent of large language models (LLMs) and Foundation Models, especially multimodal ones, presents a new frontier in ASR research. These models hold the promise of revolutionizing the field by providing more generalized and adaptable solutions.

References

[1] HUMPHREY, E. J., R. M. BITTNER, A. DEMETRIOU, S. GULATI, A. JANSSON, T. JEHAN, B. LEHNER, A. KRUSPE, A. KUMAR, S. REDDY ET AL.: *Signal processing for singing voice analysis: Significance, applications, and methods*. *IEEE Signal Processing Magazine*, 36(1), pp. 82–94, 2018.

[2] HANSEN, J. K.: *Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients*. In *Sound and Music Computing Conference (SMC)*. 2012.

[3] MAUCH, M., H. FUJIHARA, and M. GOTO: *Integrating additional chord information into hmm-based lyrics-to-audio alignment*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 20(1), pp. 200–210, 2012.

[4] KRUSPE, A. M.: *Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing*. In *17th International Society for Music Information Retrieval Conference (ISMIR)*. New York, NY, USA, 2016.

[5] ROA DABIKE, G. and J. BARKER: *Automatic lyric transcription from karaoke vocal tracks: Resources and a baseline system*. In *Interspeech*. 2019. doi:10.21437/Interspeech.2019-2378.

[6] SCHULZE-FORSTER, K., C. DOIRE, G. RICHARD, and R. BADEAU: *Phoneme level lyrics alignment and text-informed singing voice separation*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 2382–2395, 2021. doi:10.1109/TASLP.2021.3091817.

[7] BITTNER, R. M., K. PASALO, J. J. BOSCH, G. MESEGUER-BROCAL, and D. RUBINSTEIN: *vocadito: A dataset of solo vocals with f_0, note, and lyric annotations*. *CoRR*, abs/2110.05580, 2021. URL `https://arxiv.org/abs/2110.05580`. `2110.05580`.

[8] MESEGUER-BROCAL, G., A. COHEN-HADRIA, and G. PEETERS: *Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm*. In *19th International Society for Music Information Retrieval Conference*. Paris, France, 2018. Hal-02019115.

[9] MESEGUER-BROCAL, G., A. COHEN-HADRIA, and G. PEETERS: *Creating dali, a large dataset of synchronized audio, lyrics, and notes*. *Transactions of the International Society for Music Information Retrieval*, 3(1), pp. 55–67, 2020. doi:10.5334/tismir.30.

[10] HOLLIEN, H., A. R. MENDES-SCHWARTZ, and K. NIELSEN: *Perceptual confusions of high-pitched sung vowels*. *Journal of Voice*, 14(2), pp. 287–298, 2000.

[11] WANG, C.-K., R.-Y. LYU, and Y.-C. CHIANG: *An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker*. In *Interspeech*. 2003.

[12] HOSOYA, T., M. SUZUKI, A. ITO, and S. MAKINO: *Lyrics recognition from a singing voice based on finite state automaton for music information retrieval*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2005.

[13] MESAROS, A. and T. VIRTANEN: *Automatic recognition of lyrics in singing*. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, pp. 1–11, 2010.

[14] MCVICAR, M., D. P. W. ELLIS, and M. GOTO: *Leveraging repetition for improved automatic lyric transcription in popular music*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014.

[15] GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, and D. S. PALLETT: *Timit acoustic-phonetic continuous speech corpus*. In *NASA STI/Recon Technical Report N*, vol. 93. National Institute of Standards and Technology (NIST), 1993.

[16] KRUSPE, A. M.: *Training phoneme models for singing with "songified" speech data*. In *16th International Society for Music Information Retrieval Conference (ISMIR)*. Malaga, Spain, 2015.

[17] ZHANG, C., J. YU, L. CHANG, X. TAN, J. CHEN, T. QIN, and K. ZHANG: *PDAugment: Data Augmentation by Pitch and Duration Adjustments for Automatic Lyrics Transcription*. *ArXiv*, abs/2109.07940, 2021. URL `https://api.semanticscholar.org/CorpusID:237532132`.

[18] TSAI, W.-H. and H.-M. WANG: *Towards automatic identification of singing language in popular music recordings*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2004.

[19] SCHWENNINGER, J., R. BRUECKNER, D. WILLETT, and M. E. HENNECKE: *Language identification in vocal music*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2006.

[20] MEHRABANI, M. and J. H. L. HANSEN: *Language identification for singing*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011.

[21] CHANDRASKEHAR, V., M. E. SARGIN, and D. A. ROSS: *Automatic language identification in music videos with low level audio and visual features*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011.

[22] KRUSPE, A. M., J. ABESSER, and C. DITTMAR: *A gmm approach to singing language identification*. In *Proc. of the AES Conference on Semantic Audio*. London, UK, 2014.

[23] KRUSPE, A. M.: *Improving singing language identification through i-vector extraction*. In *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*. Erlangen, Germany, 2014.

[24] KRUSPE, A. M.: *Phonotactic language identification for singing*. In *Interspeech*. San Francisco, CA, USA, 2016.

[25] RENAULT, L., A. VAGLIO, and R. HENNEQUIN: *Singing language identification using a deep phonotactic approach*. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275. 2021. doi:10.1109/ICASSP39728.2021.9414203.

[26] CHOI, K. and Y. WANG: *Listen, read, and identify: Multimodal singing language identification of music*. In *International Society for Music Information Retrieval Conference*. 2021.

[27] SANTANA, I. A. P., F. PINHELLI, J. DONINI, L. CATHARIN, R. B. MANGOLIN, V. D. FELTRIM, M. A. DOMINGUES ET AL.: *Music4all: A new music database and its applications*. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 399–404. IEEE, 2020.

[28] FUJIHARA, H., M. GOTO, and J. OGATA: *Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2008.

[29] NAKANO, T. and M. GOTO: *Lyriclistplayer: A consecutive-query-by-playback interface for retrieving similar word sequences from different song lyrics*. In *Sound and Music Computing Conference (SMC)*. 2016.

[30] DITTMAR, C., P. MERCADO, H. GROSSMANN, and E. CANO: *Towards lyrics spotting in the syncglobal project*. In *3rd International Workshop on Cognitive Information Processing (CIP)*. 2012.

[31] DZHAMBAZOV, G., S. SENTÜRK, and X. SERRA: *Searching lyrical phrases in a-capella turkish makam recordings*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2015.

[32] KRUSPE, A. M.: *Keyword spotting in a-capella singing*. In *15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipei, Taiwan, 2014.

[33] KRUSPE, A. M.: *Keyword spotting in singing with duration-modeled hmms*. In *European Signal Processing Conference (EUSIPCO)*. Nice, France, 2015.

[34] LOSCOS, A., P. CANO, and J. BONADA: *Low-delay singing voice alignment to text*. In *International Computer Music Conference (ICMC)*. 1999.

[35] WANG, Y., M.-Y. KAN, T. L. NWE, A. SHENOY, and J. YIN: *Lyrically: Automatic synchronization of acoustic musical signals and textual lyrics*. In *ACM International Conference on Multimedia*. 2004.

[36] CHEN, K., S. GAO, Y. ZHU, and Q. SUN: *Popular song and lyrics synchronization and its application to music information retrieval*. In *SPIE Multimedia Computing and Networking*. 2006.

[37] MAUCH, M., H. FUJIHARA, and M. GOTO: *Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations*. In *Sound and Music Computing Conference (SMC)*. 2010.

[38] WONG, C. H., W. M. SZETO, and K. H. WONG: *Automatic lyrics alignment for cantonese popular music*. *Multimedia Systems*, 12(4-5), pp. 307–323, 2007.

[39] LEE, K. and M. CREMER: *Segmentation-based lyrics-audio alignment using dynamic programming*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2008.

[40] MESAROS, A. and T. VIRTANEN: *Automatic alignment of music audio and lyrics*. In *International Conference on Digital Audio Effects (DAFx-08)*. 2008.

[41] GONG, R., P. CUVILLIER, N. OBIN, and A. CONT: *Real-time audio-to-score alignment of singing voice based on melody and lyric information*. In *Interspeech*. 2015.

[42] DZHAMBAZOV, G., A. SRINIVASAMURTHY, S. SENTÜRK, and X. SERRA: *On the use of note onsets for improved lyrics-to-audio alignment in turkish makam music*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2016.

[43] KRUSPE, A. M.: *Lyrics alignment using hmms, posteriorgram-based dtw, and phoneme-based levenshtein alignment*. In *18th International Society for Music Information Re-*

trieval Conference (ISMIR) (MIREX submission). Suzhou, China, 2017.

[44] GUPTA, C., E. YILMAZ, and H. LI: *Acoustic modeling for automatic lyrics-to-audio alignment*. 2019. 1906.10369.

[45] STOLLER, D., S. DURAND, and S. EWERT: *End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model*. 2019. 1902.06797.

[46] VAGLIO, A., R. HENNEQUIN, M. MOUSSALLAM, G. RICHARD, and F. D'ALCHÉ BUC: *Multilingual lyrics-to-audio alignment*. In *International Society for Music Information Retrieval Conference (ISMIR)*. Montreal, Canada, 2020. URL https://hal.science/hal-02996940.

[47] DEMIREL, E., S. AHLBÄCK, and S. DIXON: *Low resource audio-to-lyrics alignment from polyphonic music recordings*. 2021. 2102.09202.

[48] SUZUKI, M., T. HOSOYA, A. ITO, and S. MAKINO: *Music information retrieval from a singing voice based on verification of recognized hypotheses*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2006.

[49] WANG, C.-C., J.-S. R. JANG, and W. WANG: *An improved query by singing/humming system using melody and lyrics information*. In *International Society for Music Information Retrieval Conference (ISMIR)*. 2010.

[50] MESAROS, A. and T. VIRTANEN: *Recognition of phonemes and words in singing*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2010.

[51] KRUSPE, A. M.: *Retrieval of textual song lyrics from sung inputs*. In *Interspeech*. San Francisco, CA, USA, 2016.

[52] KRUSPE, A. M. and M. GOTO: *Retrieval of song lyrics from sung queries*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, Canada, 2018.

[53] TSAI, C.-P., Y.-L. TUAN, and L.-S. LEE: *Transcribing lyrics from commercial song audio: the first step towards singing content processing*. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5749–5753. 2018. doi:10.1109/ICASSP.2018.8462247.

[54] DEMIREL, E., S. AHLBÄCK, and S. DIXON: *Automatic lyrics transcription using dilated convolutional neural networks with self-attention*. 2020. 2007.06486.

[55] DEMIREL, E., S. AHLBÄCK, and S. DIXON: *Mstre-net: Multistreaming acoustic modeling for automatic lyrics transcription*. 2021. 2108.02625.

[56] GAO, X., C. GUPTA, and H. LI: *Genre-conditioned acoustic models for automatic lyrics transcription of polyphonic music*. 2022. 2204.03307.

[57] OU, L., X. GU, and Y. WANG: *Transfer learning of wav2vec 2.0 for automatic lyric transcription*. 2022. 2207.09747.

[58] ZHUO, L., R. YUAN, J. PAN, Y. MA, Y. LI, G. ZHANG, S. LIU ET AL.: *LyricWhiz: Robust Multilingual Zero-Shot Lyrics Transcription by Whispering to ChatGPT*. 2023. 2306.17103.

Linguistic Politeness in Artificial Conversational Agents

Hendrik Buschmeier

Bielefeld University, Faculty of Linguistics and Literary Studies, Digital Linguistics
hbuschme@uni-bielefeld.de.de

Abstract: Politeness is a (not only) linguistic phenomenon [1] in human social interaction that manifests itself in seemingly subtle influences on the formulation of utterances and plays an important role in managing of rapport [2] and thus is crucial for communication. However, it is unclear whether and how the use of polite language transfers to interactions with artificial conversational agents such as voice assistants or social robots. Anecdotal evidence suggests that people use formulaic expression that are considered polite (such as 'please' and 'thank you') when speaking to voice assistants, and raised concerns when this was not reciprocated. In this talk, I will present politeness as a phenomenon that is much richer than formulaic speech, and talk about our own research on a computational model of politeness for conversational agents [3], empirical studies on factors influencing politeness in human-robot interaction [4], human use of indirect politeness in human-robot interaction [5], and a qualitative investigation of differences in influences on and expectations of politeness in human and human-agent interaction [6].

References

[1] Brown, P. and S. C. Levinson: *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK, 1987. doi:10.1017/CBO9780511813085.

[2] Spencer-Oatey, H.: *Face, (im)politeness and rapport*. In H. Spencer-Oatey (ed.), *Culturally Speaking. Culture, Communication and Politeness Theory*. Continuum International Publishing, 2nd edn., 2008. doi:10.5040/9781350934085.ch-002.

[3] Lumer, E. and H. Buschmeier: *Modeling social influences on indirectness in a rational speech act approach to politeness*. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, vol. 44, pp. 2796–2802. Toronto, ON, Canada, 2022.

[4] Lumer, E. and H. Buschmeier: *Perception of power and distance in human-human and human-robot role-based relations*. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 895–899. Sapporo, Hokkaido, Japan, 2022. doi:10.1109/HRI53351.2022.9889308.

[5] Lumer, E., C. Lachenmaier, S. Zarriess, and H. Buschmeier: *Indirect politeness of disconfirming answers to humans and robots*. In *Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication (Ro-Man)*, pp. 1805–1815. Busan, South Korea, 2023. doi:10.1109/RO-MAN57019.2023.10309586.

[6] Lumer, E. and H. Buschmeier: *Should robots be polite? expectations about politeness in human-robot interaction*. *Frontiers in Robotics and AI*, 10, p. 1242127, 2023. doi:10.3389/frobt.2023.1242127.