U.Porto Journal of Engineering, 8:2 (2022) 202-212
ISSN 2183-6493
DOI: 10.24840/2183-6493_008.002_0016

Received: 11 January, 2022
Accepted: 30 January, 2022
Published: 28 April, 2022

Machine Learning Approaches for Whisper to Normal Speech Conversion: A Survey

Marco A. Oliveira[1]

[1]Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 PORTO, Portugal (up199100668@edu.fe.up.pt) ORCID 0000-0002-3161-1109

Abstract

Whispered speech is a mode of speech that differs from normal speech due to the absence of a periodic component, namely the Fundamental Frequency that characterizes the pitch, among other spectral and temporal differences. Much attention has been given in recent years to the application of Machine Learning techniques for voice conversion tasks. The whisper-to-normal speech conversion is particularly challenging, however, especially with respect to the Fundamental Frequency estimation. Based on the most recent literature, this survey assesses the state-of-the-art regarding Machine Learning based whisper-to-normal speech conversion, identifying trends both on modeling and training approaches. The proposed solutions include Generative Adversarial Network based, Autoencoder based and Bidirectional Long Short-Term Memory based frameworks, among other Deep Neural Network based architectures. In addition to Parallel versus Non-Parallel training, time-alignment requirements and strategies, datasets, vocoder usage, as well as both objective and subjective evaluation metrics are also covered by the present survey.

Author Keywords. Signal Processing, Machine Learning, Whispered Speech, Normal Speech, Voice Conversion, Speech Conversion.

Type: Review Article
Open Access Peer Reviewed CC BY

1. Introduction

Whispered speech is a special mode of speech that differs from normal speech, most noticeably for the lack of vocal folds contribution during speech production, that is to say that it lacks phonation (or voicing). Since the vocal folds do not vibrate during whispered speech, the resulting speech signal lacks the periodic component that is present in certain regions of normal speech, namely during vowels and voiced regions of voiced consonants. Hence, this signal has noisier characteristics, tending to be less intelligible and more susceptible to the interference from surrounding noise sources (Silva, Oliveira, and Ferreira 2021). Despite these disadvantages, individuals may still communicate through whisper, either intentionally (*e.g.*, due to privacy reasons) or as result of a temporary or permanent condition, such as vocal folds paralysis or lack of vocal folds due to a laryngectomy, among others (Lian et al. 2019a).

In a broad sense, Voice Conversion (VC) systems aim at altering speech characteristics while preserving the original signal linguistic content (Huang, Lin, and Lee 2021). Its applications range from manipulating speaker/gender identity, style *(e.g.*, from neutral to emotional speech) or mode of speech, such as from normal speech to singing, Lombard or whispered speech. Such systems have been greatly improved in recent years through the application of Machine Learning (ML) techniques, including Deep Neural Networks, or DNNs (Huang, Lin, and Lee 2021). The conversion from whispered to normal speech, however, is particularly

challenging largely due to the absence of a Fundamental Frequency (F_0) present in the original signal to rely on during reconstruction, as a result of the aforementioned lack of vocal folds contribution (Parmar et al. 2019). To the best knowledge of the author of this survey, presently, there is still no efficient method of reliably estimating the pitch of the targeted normal speech from the remaining acoustic cues present in the original whispered speech. Furthermore, while a well estimated F_0 is of the most importance due to its perceptual impact in terms of prosody and naturalness (Silva, Oliveira, and Ferreira 2021), there are other spectral and temporal differences between whisper and normal speech that conversion systems may need to address as well (Parmar et al. 2019). Motivations for an efficient whispered to normal speech conversion may include allowing patients to communicate in a more natural and comfortable manner or to improve human-machine interaction in the context of Automatic Speech Recognition (ASR), especially given the growing prominence of voice assistants in recent years (Niranjan et al. 2020).

The present survey aims mainly at identifying which ML approaches have been adopted in the most recent literature w.r.t whisper-to-normal speech conversion, discussing not only what the main challenges are but also how they have been addressed in the most recent proposals. Additional objectives include identifying which vocoders, datasets and evaluation metrics are most commonly used in this context, as well. The remaining of this document is organized as follows: Sec.2 addresses the main differences between whispered and normal speech from the speech production and signal processing perspectives and how they are typically approached by VC systems; Sec.3 presents the surveying methodology and the resulting paper selection; Sec.4 describes the respective modelling and training approaches; Sec.5 covers the dataset and vocoder choices; Sec.6 addresses the evaluation metrics implemented in the original papers; and lastly, Sec.7 concludes with a brief discussion and final remarks.

2. Whispered *vs.* Normal Speech

As previously indicated, whispered speech is produced without the contribution of the vocal folds. In this case, the speech signal is the result of constricted turbulent air that flows through the glottis (excitation) and that is further modulated by the oral and nasal cavities (articulation) that composes the vocal tract (Perrotin and McLoughlin 2020). As a result of the different coupling between the trachea and the vocal tract, the whispered speech signal tends to differ from normal speech in terms of formant shape and location, most noticeably in the F1 and F2 frequencies region, and in terms of spectral slope as well (Parmar et al. 2019). Also, during whispered speech the lungs need to exhale more airflow compared to normal speech (Wolfe, Garnier, and Smith 2009) and to compensate for this, phoneme duration tends to be longer than that of their normal speech counterparts (Gao et al. 2021). Such temporal differences, which are especially important under parallel training scenarios, are often addressed with pre-alignment via Dynamic Time Warping (DTW). When used during the training phase, however, this approach is noted for its propensity for resulting in artifacts, compromising both the intelligibility and the perceptual quality of the converted speech (Gao et al. 2021; Lian et al. 2019a). As for the spectral differences and the lack of F_0, the typical approach consists in starting by remapping the spectral envelope from whispered to its normal speech counterpart, for which there are several approaches and often involves converting the spectral envelope into Mel cepstral features, such as Mel Frequency Cepstral Coefficients (MFCC). This is usually followed by predicting the F_0 contour based on the original whispered features and/or the converted ones. Such approach is guided partly by the fact that while F_0 is absent in the original whispered speech, a sensation of pitch has been observed in the literature. However, the relation between spectral envelope and F_0 is rather intricate and

the F_0 contour obtained by such methods has limited accuracy, compromising the naturalness of the converted normal speech (Parmar et al. 2019). Arguably, the estimation of the F_0 remains the most challenging task in whisper-to-normal speech conversion.

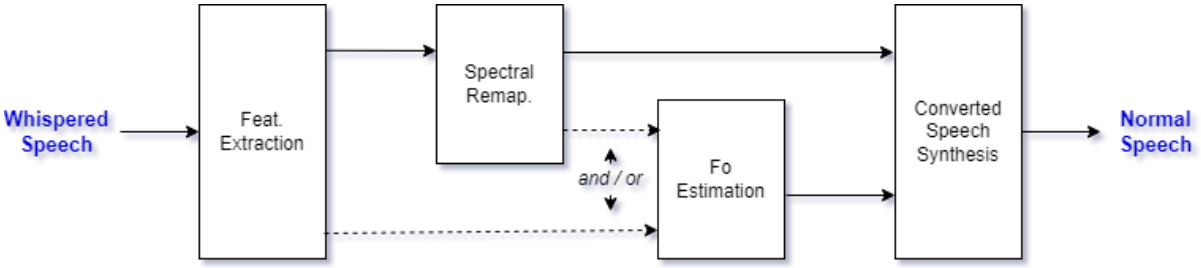

Figure 1: Illustration of the whisper-to-normal speech typical framework pipeline

The diagram presented in Figure 1 illustrates the typical whisper-to-normal speech conversion framework pipeline, as was described above. As it will be shown in Sec.4, however, how these subtasks are implemented and interact with each other may vary.

3. Paper Selection

The relevant papers were surveyed through Scopus, Web of Science and Google Scholar. Given the rapid evolution of the Machine Learning landscape and the aim at identifying current trends and challenges, only the most recently published works, namely from 2019 onwards, were included. Required terms either in the Title, Abstract or Keyword fields included: whisper OR whispered, normal OR neutral OR natural, speech OR voice OR conversion OR converting. In addition to the prerequisite of adopting ML approaches for at least one of the conversion subtasks, it was also required that the conversion systems made explicit usage of natural whispered speech as source domain and having synthetic normal speech as target domain. Hence, works aiming at producing synthetic whispered speech and/or having a different source domain (*e.g.*, Text-to-Speech systems) were excluded. A total of 8 papers were collected via Scopus, namely Lian et al. (2019a), Patel et al. (2021), Parmar et al. (2019), Malaviya et al. (2020), Pang et al. (2020), Yu et al. (2019), Lian et al. (2019b), and Lian et al. (2020). One of these, namely Lian et al. (2020), was excluded from the survey due to the lack of availability of a full text version. Additionally, following the same criteria, 3 other papers were collected through Web of Science and/or Google Scholar, namely Niranjan et al. (2020), Gao et al. (2021), and Patel et al. (2019). It is also worth noting that, although beyond the scope of this survey, the prior research in whisper-to-normal speech conversion is largely covered by the background sections within the selected papers.

4. Modelling and Training Approaches

This section does not intend to present a detailed description of each conversion system, nor the architecture details behind them, which may be found via the original papers. The focus is put instead on the respective model and training approaches and how the key challenges of converting whispered-to-normal speech are being addressed. In most cases, the systems will be referred by names given by the respective authors (note that in two cases, an identifiable name was used for conveniency just for the propose of this survey).

4.1. DNN-MCC-F0

In Niranjan et al. (2020), inspired by the results of Deep Neural Networks in diverse visual applications, two distinct 5-layer DNN were trained for the spectrum features remapping and for the F_0 estimation subtasks, respectively. The first one uses the Mel Cepstral Coefficients (MCC) from whisper and from normal speech, which are aligned through DTW, as to learn their relationship. To reflect the dynamic characteristics of the spectral envelope, the first

order differences of the MCC were also included. The second one uses the MCC both from whisper and normal speech and the F_0 from the reference normal speech so to learn to predict the F_0 of the converted speech. Parallel training was conducted with 300 utterances both in whisper and normal speech counterparts. The authors reported better results when also including MCC deltas in addition to the basic MCC values, and better F_0 prediction than with a previously proposed Gaussian Mixed Model (GMM) based method, available in the literature.

4.2. SEQ2SEQ

In Lian et al. (2019a), the authors adopted the SEQ2SEQ mapping framework, firstly proposed by Cho et al. (2014), to map the relationship between the MFCCs of the whispered speech and the correspondent normal speech. An auditory attention mechanism was also adopted as to obtain a self-adaptive context vector that is used to adaptively estimate the current hidden state and output of the decoder. Through this method, the requirement of pre-alignment before training the parallel whisper and normal speech is avoided. Training was conducted through 300 parallel whisper and normal speech utterances, plus 48 parallel utterances for testing. The SEQ2SEQ remaps 30-dimensional MFCC vectors, that were obtained from spectral representations of each frame, from whisper to normal speech. Once the MFCCs of the estimated normal speech are obtained, these are used to train two Deep Bidirectional LSTM (Long Short-Term Memory), or DBLSTM, that are responsible for characterizing the relationship between the estimated MFCCs and the F_0 of the normal speech and for estimating the signal aperiodic component, respectively. The authors argue that their method outperforms the classical DTW approach, reporting better results both in terms of naturalness and F_0 prediction than those of alternative methods.

4.3. AGAN-W2SC

In Gao et al. (2021), a Generative Adversarial Network (Goodfellow et al. 2014), or GAN, is responsible for remapping the whispered features to the target domain of normal speech. The generator part of the GAN, an encoder-decoder, is implemented through a Siamese pair of convolutional networks. The discriminator part is composed by a single convolutional network. The authors argue that the system can implicitly generate the fundamental frequency without explicitly trying to predict it (as it is traditionally done), although they are vague with the details. Also, in a similar approach to the previous system, the authors argue that the requirement of pre-alignment before training is circumvented by an adaptive time-alignment through the inclusion of a self-attention mechanism in the encoder. In this case, this is achieved by processing pivotal features and assigning weights to each region adaptively so to implicitly perform the time alignment. The system was trained with 800 pairs of parallel corpuses (plus 169 other pairs reserved for testing) from which were extracted frame-level Mel-spectrogram vectors. Better results were reported in terms of perceptual quality and intelligibility and competitive ones in terms of F_0 prediction compared to DTW based methods.

4.4. Inception-GAN vs. CNN-GAN

In Patel et al. (2019), the authors adapted the Inception modules from Szegedy et al. (2015), proposing an Inception-GAN architecture, aimed at reducing computational complexity. In contrast with a typical CNN based GAN, also implemented as baseline for comparison purposes, the generator and discriminator parts of the GAN are made of stacks of 4 and 3 inception modules, respectively, and only 1 convolutional layer each. Following this alternative architecture, one model was used for remapping the cepstral features (MCC) from the whispered to the normal speech and another one to find the prediction function from the converted features to the F_0 of the targeted normal speech. This was followed by smoothing

the voiced regions in post-processing. The authors used the traditional GAN training approach, conducting parallel training with 1164 utterances, both in whispered and normal speech modes, plus 35 additional utterances for testing. Alignment is only mentioned during the evaluation phase, which was carried out via DTW, as to measure the F_0 prediction accuracy. Better results were reported in terms of naturalness for the Inception based architecture compared to the CNN based baseline with speaker-specific conversions, both for a male and a female speaker.

4.5. CycleGAN vs. DiscoGAN

In Parmar et al. (2019), the authors adopted both the CycleGAN (Zhu et al. 2017) and the DiscoGAN (Kim et al. 2017) architectures as to evaluate their effectiveness for whisper-to-normal speech conversion. These architectures, originally introduced independently for unpaired image-to-image translation and since then adapted for audio conversion as well, although following the same modelling and training approach, differ with each other w.r.t the respective objective functions. In both approaches, one model was implemented with the purpose of remapping the cepstral features from the whispered to the normal speech and another one to map from the converted cepstral features (MCC) to the corresponding F_0 of normal speech, which was followed once again by smoothing the voiced regions in post-processing. The two models are trained sequentially in this order. It is also indicated that the voiced-unvoiced decision (*i.e.*, which regions of the reconstructed signal are to be voiced) is implemented in both architectures through the same DNN-based model, without providing further details about this module. According to the authors, 388 parallel utterances corresponding to whispered and normal speech were used for training, plus 35 utterances for testing. Alignment through DTW is also only mentioned during testing phase, in the context of measuring F_0 accuracy, implying that there was no pre-alignment before training. According to the authors, both architectures outperformed the traditional GANs in terms of naturalness.

4.6. CinC-GAN

The authors argue that the previously mentioned CycleGAN architecture led to additional non-linear noise in the predicted F_0, due to this prediction being highly dependent on the pre-trained MCC mapping (*v.s.*, CycleGAN sequential training). In Patel et al. (2021), to improve the effectiveness of F_0 prediction without sacrificing the accuracy of the MCC mapping, a new architecture was proposed that instead of relying on two separate models that are trained sequentially, relies on a single model containing an inner CycleGAN for MCC mapping and an outer CycleGAN for F_0 prediction, hence jointly training this sub-networks. Also, as pointed out by the authors, non-parallel training was conducted in all experiments which means that there was no need for time-alignment. These experiments included speaker-specific, gender-specific, and both seen and unseen speakers, reporting better results compared to the CycleGAN baseline, both in terms of naturalness and F_0 prediction. The authors highlighted the better performance obtained by the new architecture with unseen speakers.

4.7. Mspec-Net

In Malaviya et al. (2020), a multi-domain speech conversion system is proposed, capable of converting both from Non-Audible Murmur (NAM) and from whispered speech to normal speech, through three domain-specific AutoEncoders (AEs). These AEs are used to obtain an internal representation of features, which are known as latent representations. During training, the aim is to learn a common latent space of the input speech from all the three domains. During conversion, only the source encoder and the target decoder are used. Like in the three previous systems, for the spectral remapping, MCC were extracted from the source

domain signals and remapped to the target domain of normal speech. For the F_0 prediction task, however, the previously mentioned CycleGAN was used instead. Prior to training, DTW was used to align the normal and whispered speech. NAM and whispered speech were recorded simultaneously, hence no further alignment was needed for this pair. With respect to whisper-to-normal conversion, the authors reported better results in terms of naturalness with Mspec-Net compared to the DiscoGAN baseline.

4.8. Meta-BLSTM

In Yu et al. (2019), a RNN (Recurrent Neural Network) based Bi-directional LSTM (BLSTM) approach is used for the remapping tasks. BLSTM based converters are known to produce high quality conversions in terms of naturalness but tend to suffer from model complexity and inference cost (Nisha Meenakshi and Ghosh 2018). Aiming at reducing complexity, the authors proposed a meta-network with non-shared weights for the LSTM memory block, reducing the number of parameters compared to the standard BLSTM. In their approach, three models were trained to map the relationship between the MCC extracted from the whispered speech and the MCC, the aperiodic component and the F_0 of the converted normal speech, respectively. Parallel training was conducted with 300 pre-aligned utterances in whisper and normal speech, plus 37 utterances for testing. Pre-alignment was implemented via DTW. The authors argue that the novel Meta-BSLTM achieved state-of-the-art results, comparable and slightly better than the baseline BLSTM, while drastically reducing the training time.

4.9. LSTM-SP+

In Pang et al. (2020), a RNN based BLSTM is used again, following the typical approach of remapping the spectral features from the whisper speech to the target normal speech, followed by F_0 estimation based on the obtained converted features. However, aiming at improving F_0 prediction, an additional 76-dimensional input is used through a feature fusing process that includes both MFCCs and prosody related features. According to the authors, these prosody related features include formants information, several energy related features and the short-term average zero-crossing rate. The system was trained on a frame-by-frame basis using a parallel corpus of 348 utterances in whispered and corresponding normal speech. The authors reported improved speech quality and better F_0 prediction when those supplemental fused features were used, compared to the baseline.

4.10. E2E-W2NSC

The speech conversion is implemented in Niranjan et al. (2020) having machine intelligibility in mind. Unlike the previously mentioned works, this conversion relies on an end-to-end system: a custom DNN comprising an encoder and a decoder, with similar architectures, both with self-attention mechanisms in addiction to feedforward fully connected layers, trained as a whole. Since the actual objective in this case is to improve ASR, which is typically trained on and optimized for normal speech recognition, the lack of F_0 in the original whispered speech is not addressed during conversion. The authors focused their attention on the formant location aspect instead, namely from F1 to F4. The network uses frame-wise acoustic features, either MCC or smoothed spectral features as input, with both alternatives being evaluated by the authors, and outputs the correspondent features for the targeted natural speech. The proposed system can use both parallel and non-parallel data. For the parallel training, which requires matched time-duration segments, trimming was conducted, with additional time-stretching to account for the remaining marginal differences. No other alignment procedure is mentioned. The authors reported better results with the converted speech compared to the ASR with the original whispered speech in a variety of experiments. The models were also pre-

trained with larger amounts of non-parallel data, as to address the sparsity problem of whispered/normal speech parallel data, which was found to improve system performance.

5. Datasets and Vocoders

This section provides a brief description of the datasets and vocoders used in the papers included in the present survey.

5.1. Datasets

Some version of a dataset derived from the TIMIT corpus was used in all cases included here. Specifically, whispered TIMIT (wTimit)[1] was used by Niranjan et al. (2020), Patel et al. (2021), Parmar et al. (2019), and Patel et al. (2019), while CSTR-NAM-TIMIT Plus was used by Gao et al. (2021), Lian et al. (2019a), Malaviya et al. (2020), Pang et al. (2020), Yu et al. (2019), and Lian et al. (2019b). The wTIMIT dataset uses the prompts in TIMIT, a well-known corpus often used for benchmarking in speech recognition, including 450 phonetically balanced sentences both in normal and whispered speech. The corpus includes Singaporean English and North American English accents. The CSTR-NAM-TIMIT Plus[2] dataset is another TIMIT derived corpus with both whispered and NAM recordings, including 421 sentences selected from newspaper text, 460 sentences from TIMIT and 18 isolated words. Some of the papers expanded the dataset with custom recordings to better suit the specific requirements of their systems. Also, in Niranjan et al. (2020) SpeechOcean was combined for training with wTIMIT in addition to custom recordings, and CHAINS[3] and LibriSpeech[4] were also used for testing and additional experiments.

5.2. Vocoders

A parametric vocoder was used in all cases presented in this survey, either for feature extraction (analysis) and/or signal reconstruction (synthesis). STRAIGHT5 was used by Gao et al. (2021), Lian et al. (2019a), Pang et al. (2020), Yu et al. (2019), and Lian et al. (2019b), WORLD6 is used by Niranjan et al. (2020) and AHOCODER7 is used by Malaviya et al. (2020), Parmar et al. (2019), Patel et al. (2021), and Patel et al. (2019). A neural vocoder, WaveNet, for higher quality synthesis, is proposed in Parmar et al. (2019) and Patel et al. (2019), but only as future work. STRAIGHT is a widely used parametric vocoder, available in several versions and with analysis and synthesis capabilities, that decomposes the speech signal and extracts the F0 contour, a frame based spectral representation and the aperiodic component allowing for their manipulation and posterior signal reconstruction. WORLD is a publicly available vocoder that follows the same architecture and extracts those same components, differing only on the algorithmic implementation. AHOCODER on the other hand, while having similar capabilities, extracts different features, namely the log(F0), a cepstral representation of the spectral envelope and the maximum voiced frequency. Nevertheless, all these vocoders are similarly adequate for speech manipulation or voice conversion.

---

[1] http://www.isle.illinois.edu/sst/data/wTIMIT
[2] https://datashare.ed.ac.uk/handle/10283/3849
[3] https://chains.ucd.ie
[4] https://www.openslr.org/12
[5] https://github.com/HidekiKawahara/legacy_STRAIGHT
[6] https://github.com/mmorise/World
[7] https://aholab.ehu.eus/ahocoder/info.html

6. Evaluation

This section discusses the evaluation metrics, both of objective and subjective natures, that were used within the surveyed papers. Code and demo samples availability is also discussed.

6.1. Objective metrics

The most used objective metric is the Mel Cepstral Distortion (either referred as CD or MCD), which was included in all papers except for Niranjan et al. (2020) and Gao et al. (2021). The MCD is given by:

$$MCD = \frac{10}{log10}\sqrt{2\sum_{d=1}^{D}(C_d - C'_d)^2} \quad (1)$$

where C_d and C'_d represent the *dth* element of the cepstral coefficients feature of the reference and the converted normal speech, respectively, and D represents the dimension of the cepstral coefficient feature. A higher MCD value indicates a greater difference between the converted and the reference speech (lower is better). Additionally, Short-Time Objective Intelligibility (STOI) was also used in Lian et al. (2019a), Pang et al. (2020), Yu et al. (2019), and Lian et al. (2019b) and Perceptual Evaluation of Speech Quality was also used in Lian et al. (2019a), Pang et al. (2020), and Yu et al. (2019). Finally, the two remaining papers used entirely different metrics from the rest. In Gao et al. (2021), Single Sided Speech Quality Assessment (P.563) was used instead. In Niranjan et al. (2020), which is automatic speech recognition oriented, Word Error Rate (WER) and Bilingual Evaluation Understudy (BLEU) were the objective metrics implemented to evaluate system performance. This paper also proposed and implemented Formant Divergence Metric (FDM) to compare formant distribution between converted and natural speech.

With respect to the accuracy of the F_0 estimation, Root Mean Square Error (RMSE) is the widely adopted metric, which is given by:

$$RMSE(F_0) = \sqrt{\sum_{i=1}^{K}\left(F_o^C{}_i - F_o^R{}_i\right)^2} \quad (2)$$

where F_0^C and F_0^R are the fundamental frequencies of each of the K time-aligned frames of the converted and the reference normal speech, respectively. This metric was adopted in Gao et al. (2021) and Lian et al. (2019a) and in all the AHOCODER based works, namely Parmar et al. (2019), Malaviya et al. (2020), Patel et al. (2021), and Patel et al. (2019), in which cases the RMSE was calculated over the log(F_0) instead. Additionally, in Patel et al. (2021), the Kullback-Leibler Divergence (KLD) and Jensen-Shannon Divergence (JSD) between predicted and original F_0 for speaker-specific tasks were also considered. In Niranjan et al. (2020), F_0 is not estimated and in Pang et al. (2020), Yu et al. (2019), and Lian et al. (2019b), while the F_0 estimation is discussed and accompanied with illustrations of the F_0 curve, no objective metric is presented.

6.2. Subjective metrics

While all papers supported their results with at least one objective metric, only a few provided a subjective evaluation, namely: Lian et al. (2019a), Parmar et al. (2019), Patel et al. (2021), Malaviya et al. (2020), and Patel et al. (2019). The lack of a subjective evaluation is justified in Niranjan et al. (2020) since the VC was implemented in the context of ASR and having machine intelligibility in mind. Subjective evaluations, since carried by human subjects, tend to be costly and time consuming compared to objective evaluations, but they are of the most

importance every time humans are intended as the end user of the system. Even more so because humans not always agree with objective results and systems that perform better with such metrics may introduce artifacts that are rather disruptive from a perceptual standpoint, as noticed by human subjects. So, in such a context, without a subjective evaluation, it will remain unclear if there is a real advantage even if the system performs better with certain objective metrics. Regarding the papers that included subjective evaluation metrics, Mean Opinion Score (MOS) was used in Lian et al. (2019a), Patel et al. (2021), Malaviya et al. (2020), and Patel et al. (2019) to assess the naturalness of the converted speech, with the number of subjects ranging from 16 up to 28. An ABX preference test between the proposed and a reference system was also conducted in Lian et al. (2019a) and Parmar et al. (2019).

System	Training	MCD (lower is better)	MOS (higher is better)
SEQ2SEQ	P	2,8	3,6
CinC-GAN	NP	6,3	3,7
CycleGAN	P/NP	6,6	-
DiscoGAN	P/NP	6,6	-
Mspec-Net	P	3,0	4,5
Inception-GAN	P	6,7	2,8
CNN-GAN	P	7,2	2,4
BLSTM-SP+	P	4,9	-
Meta-BLSTM	P	4,8	-
DNN-MCC-F0	P	5,8	-

Table 1: Summary presenting the most used objective and subjective metrics, MCD and MOS respectively, where available. The training approach, parallel (P) or non-parallel (NP) is also indicated. These results are not intended as providing a direct comparison between the included systems, tested under different conditions

Table 1 summarizes the most used objective and subjective metrics, respectively MCD and MOS, where available. In most of these cases, the systems were tested under different conditions and evaluated by different subjects. Hence, the table is not intended as a direct comparison between these systems, but merely as indicative. For convenience, results were averaged down in some of the cases (*e.g.*, if a result was provided for male speakers and another for female speakers in the original paper).

6.3. Code and demo samples availability

While not providing an evaluation by itself, demo samples are useful for any interested reviewer or fellow researcher as an indication of the capabilities and level of quality attained by each conversion system. Among the papers included in this survey, however, only two provide a webpage with demo samples, namely Gao et al. (2021) and Malaviya et al. (2020). Note that several other systems (some of which are covered by this survey) were also included in those two sets, albeit not always by their own original authors. While conducting subjective tests is a consuming task, providing demos that would help to put the objective results in context, would be a relatively simple task and in this author opinion, should be encouraged. Code is also made available via GitHub in Niranjan et al. (2020), Gao et al. (2021), Malaviya et al. (2020), allowing for the possibility of replication or other experiments for a limited number of the surveyed papers.

7. Conclusions

This survey researched the most recent literature regarding ML based whisper-to-normal speech conversion. A total of 10 papers were included, discussing their modelling and training approaches. These include GAN based model approaches in 4 papers, one AE based model

and two BLSTM based models, among other custom DNN designs. Most of these implementations relied in parallel training, making use, and eventually expanding upon one of two publicly available TIMIT derived datasets, namely wTIMIT and CSTR-NAM-TIMIT Plus. One paper only focused specifically on non-parallel training and another one used non-parallel pre-training followed by fine-tuning with parallel data, as a data augmentation strategy. Regarding time-alignment, only 3 papers indicated applying pre-alignment via DTW, while two other papers reported using DTW based time-alignment in the testing phase in order to measure F_0 estimation accuracy. Also, all papers included made use of a parametric vocoder to extract the basic whisper and normal speech features and/or reconstructing the converted speech with one paper indicating plans of future application of a neural vocoder.

Regarding the reported results, all papers made use of at least one objective metric to support their claims, most often measuring the Mel Cepstral Distortion between the features of the converted speech and those of the reference normal speech, as well as RMSE for F_0 prediction accuracy. However, only 5 papers included subjective tests: 4 papers included MOS based perceptual tests to assess the naturalness of the converted speech and ABS preference tests between the proposed a reference system were conducted in 2 papers (note that one paper included both). Although most papers reported F_0 estimation accuracy improvements through the proposed methods compared to the baseline, authors tend treat the F_0 estimation as an open problem and going as far as admitting that better F_0 estimation is still required. Also, the number of papers that focused on non-parallel learning is rather limited. Whenever parallel learning is required, it represents an important limitation, tending to be costly and time consuming, hence not being an attractive and practical solution for a system intended for real world usage. The availability of demo samples or code, as to better assess the proposed systems capabilities, was shown to be limited among the surveyed papers as well.

References

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". Preprint, submitted June 13, 2014. https://arxiv.org/abs/1406.1078.

Gao, T., J. Zhou, H. Wang, L. Tao, and H. K. Kwan. 2021. "Attention-guided generative adversarial network for whisper to normal speech conversion". Preprint, submitted November 2, 2021. https://arxiv.org/abs/2111.01342.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative adversarial networks". Preprint, submitted June 10, 2014. https://arxiv.org/abs/1406.2661.

Huang, T. H., J. H. Lin, and H. Y. Lee. 2021. "How far are we from robust voice conversion: A survey". In *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, 514-21. IEEE. https://doi.org/10.1109/SLT48900.2021.9383498.

Kim, T., M. Cha, H. Kim, J. K. Lee, and J. Kim. 2017. "Learning to discover cross-domain relations with generative adversarial networks". Preprint, submitted March 15, 2017. https://arxiv.org/abs/1703.05192.

Lian, H., Y. Hu, W. Yu, J. Zhou, and W. Zheng. 2019a. "Whisper to normal speech conversion using sequence-to-sequence mapping model with auditory attention". *IEEE Access* 7: 130495-504. https://doi.org/10.1109/ACCESS.2019.2940700.

Lian, H., Y. Hu, J. Zhou, H. Wang, and L. Tao. 2019b. "Whisper to normal speech based on deep neural networks with MCC and F0 features". In *International Conference on Digital Signal Processing, DSP*. IEEE. https://doi.org/10.1109/ICDSP.2018.8631888.

Lian, H., J. Zhou, Y. Hu, and W. Zheng. 2020. "Whisper to normal speech conversion using deep convolutional neural networks". *Shengxue Xuebao/Acta Acustica* 45, no. 1: 137-44.

Malaviya, H., J. Shah, M. Patel, J. Munshi, and H. A. Patil. 2020. "Mspec-Net: Multi-domain speech conversion network". In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 7764-68. IEEE. https://doi.org/10.1109/ICASSP40776.2020.9052966.

Niranjan, A., M. Sharma, S. B. C. Gutha, and M. A. B. Shaik. 2020. "End-to-end whisper to natural speech conversion using modified transformer network". Preprint, submitted April 20, 2020. https://arxiv.org/abs/2004.09347.

Nisha Meenakshi, G., and P. K. Ghosh. 2018. "Whispered speech to neutral speech conversion using bidirectional LSTMs". In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 491-95. https://doi.org/10.21437/Interspeech.2018-1487.

Pang, C., H. Lian, J. Zhou, H. Wang, and L. Tao. 2020. "Method for transforming whisper to normal speech with feature fusion". *Nanjing Hangkong Hangtian Daxue Xuebao/Journal of Nanjing University of Aeronautics and Astronautics* 52, no. 5: 777-82. https://doi.org/10.16356/j.1005-2615.2020.05.014.

Parmar, M., S. Doshi, N. J. Shah, M. Patel, and H. A. Patil. 2019. "Effectiveness of cross-domain architectures for whisper-to-normal speech conversion". In *European Signal Processing Conference*. IEEE. https://doi.org/10.23919/EUSIPCO.2019.8902961.

Patel, M., M. Parmar, S. Doshi, N. Shah, and H. A Patil. 2019. "Novel Inception-GAN for Whisper-to-Normal speech conversion". In *Proceedings 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 87-92. https://doi.org/10.21437/SSW.2019-16.

Patel, M., M. Purohit, J. Shah, and H. A. Patil. 2021. "CinC-GAN for effective F0 prediction for whisper-to-normal speech conversion". In *European Signal Processing Conference*, 411-15. IEEE. https://doi.org/10.23919/Eusipco47968.2020.9287385.

Perrotin, O., and I. V. McLoughlin. 2020. "Glottal flow synthesis for whisper-to-speech conversion". *IEEE/ACM Transactions on Audio Speech and Language Processing* 28: 889-900. https://doi.org/10.1109/TASLP.2020.2971417.

Silva, J., M. Oliveira, and A. Ferreira. 2021. "Flexible parametric implantation of voicing in whispered speech under scarce training data". In *European Signal Processing Conference*, 416-20. IEEE. https://doi.org/10.23919/Eusipco47968.2020.9287684.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2015. "Rethinking the nception architecture for computer vision". Preprint, submitted December 2, 2015. https://arxiv.org/abs/1512.00567.

Wolfe, J., M. Garnier, and J. Smith. 2009. "Vocal tract resonances in speech, singing, and playing musical instruments". *HFSP Journal* 3, no. 1: 6-23. https://doi.org/10.2976/1.2998482.

Yu, W., H. Lian, J. Zhou, H. Wang, and L. Tao. 2019. "Whispered speech to normal speech conversion using bidirectional LSTMs with meta-network". In *2019 2nd IEEE International Conference on Information Communication and Signal Processing, ICICSP 2019*, 251-55. IEEE. https://doi.org/10.1109/ICICSP48821.2019.8958537.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros. 2017. "Unpaired image-to-image translation using cycle-consistent adversarial networks". Preprint, submitted March 30, 2017. https://arxiv.org/abs/1703.10593.