Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion

Luc Ardaillon, Nathalie Henrich Bernardoni, Olivier Perrotin
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
{luc.ardaillon, nathalie.henrich, olivier.perrotin}@gipsa-lab.fr

Abstract
Corpectomized or laryngectomized patients recover the ability to speak thanks to devices able to produce a natural-sounding voice source in real time. However, constant voicing can impair the naturalness and intelligibility of reconstructed speech. Voicing decision, consisting in identifying whether an uttered phone should be voiced or not, is investigated here as an automatic process in the context of whisper-to-speech (W2S) conversion systems. Whereas state-of-the-art approaches apply DNN techniques on high-dimensional acoustic features, we seek here a low-resource alternative approach for a perceptually-meaningful mapping between acoustic features and voicing decision, suitable for real-time applications. Our method first classifies whisper signal frames into phoneme classes based on their spectral centroid and spread, and then discriminates voiced phonemes from their unvoiced counterpart based on class-dependent spectral centroid thresholds. We compared our method to a simpler approach using a single centroid threshold on several databases of annotated whispers in both single-speaker and multi-speaker training setups. While both approaches reach voicing accuracy higher than 91%, the proposed method allows to avoid some systematic voicing decision errors, which may allow users to learn to adapt their speech in real-time to compensate for remaining voicing errors.

Index Terms: Whisper-To-Speech conversion, spectral moments, voiced/unvoiced decision, phonemes classification

1. Introduction
Voice disorders inducing a loss of voicing (e.g. following a cordectomy or laryngectomy) severely hamper the patient’s communicative abilities due to the loss of vocal source information such as pitch contour and voicing decision [1]. The latter is essential for intelligibility since speech normally alternates between phonated vowels and phonated or unphonated consonants. Loss of voicing information introduces ambiguities among consonants that all become unphonated, as in whisper [2]. One of the main challenges of voice-source rehabilitation systems is to accurately predict and introduce the missing vocal source information. For this purpose, non-invasive solutions like electrolarynx [3, 4] generate an artificial voiced signal that emulates a glottal source from which the user can articulate voiced speech. Such system imposes a constant voicing while articulating, which may thus limit naturalness and intelligibility of unvoiced segments. Solutions to activate voicing manually based on finger or arm movements were proposed [5, 6], yet not suitable to control rapid alternation between voiced and unvoiced segments that occur in normal speech. Therefore, automatic signal-based approaches to voicing decision have been broadly adopted in recent systems, many of which tackling whisper-to-speech (W2S) conversion tasks [7, 8, 9, 10, 11, 12, 13]. Since the ability to whisper does not require functional vocal folds, W2S is a valid paradigm for voice-source reconstruction. All these recent methods use extensive deep neural networks (DNN) for voicing decision, tested in offline configurations. Apart from the current challenge of embedding complex neural networks on real-time processors, we believe that, given the proximity of voiced and unvoiced consonants with a similar place of articulation when they are whispered [2], the voicing decision may asymptotically improve but hardly be perfect. Alternatively, [14, 15] demonstrated the ability of speakers to adapt their own speech production when their auditory feedbacks are altered by a real-time voice modification system. Therefore, providing that the mapping between acoustic features and voicing decision in W2S systems is perceptually meaningful, we hypothesise that the speaker can learn this mapping and adapt his/her acoustic features to compensate for the remaining voiced decision errors. For instance, hyper-articulation has been discussed as a way to better discriminate voiced consonants from their unvoiced counterparts [16]. In this line, direct inference of a voicing decision from the spectral centre of gravity (or centroid) of whisper showed promising results [17], but the use of one single decision threshold for all phonemes resulted in the systematic miscategorisation of some phonemes. It has been proposed to first classify signal frames into phoneme classes using pre-defined rules based on temporal or frequency-band energy variation [18, 19], but their voicing decisions based on manually-set and speaker-dependant thresholds lack robustness and call for more generalising machine-learning approaches. To avoid falling back in the need to acquire large quantity of annotated data as DNN often require, that is rare for whispered speech and usually not well-suited for real-time applications, it is important to target small-size machine-learning algorithms while paying particular care in designing the training corpus.

Therefore, we propose a novel whispered-phoneme voicing classifier based on a 2-step procedure that: 1) classifies signal frames into naturally voiced phonemes and three types of fricatives. A simple KNN method is investigated and several training corpora varying in size, content and annotation are tested. 2) subdivides each fricative class into voiced and unvoiced using a simple centroid threshold. Voicing decision results will be discussed regarding the potential of error compensation with input adaptation offered by the system, but the real-time implementation and human control evaluation is out of scope of this paper. In the following, the proposed approach is presented in section 2. Section 3 describes the data and conditions used in our evaluations. Section 4 presents the results of this study.

2. Proposed approach
2.1. Choice of descriptors
Most recent machine learning-based studies that are trained on parallel corpus of normal and whispered speech predict
voicing decision from Mel-Frequency Cepstral Coefficients, a high-dimensional representation of the whisper spectral envelope [7, 20, 21, 10, 11, 22]. Although they reach low voicing decision errors rates (6.8% and 5% for [7, 11] respectively), the input features are computed over relatively long audio segments which is not well suited for real-time. Their high dimensionality along with the non-linear mapping introduced by the neural network make the process difficult to understand for a user to be able to anticipate and compensate for the remaining voicing decision errors. For this sake, it is preferable to use only few identified salient features to perform classification. Energy ratios between different frequency bands can be used to classify phonemes [18, 19], but their relevance may vary depending on the place of articulation of the fricative (and thus on co-articulation). Computing spectral moments seems a more promising approach in globally considering the spectrum as a statistical distribution [23]. The two first moments, centroid and spread (or standard deviation), were found to be of particular interest for discriminating the three different types of French fricatives [23]. We thus investigate the use of these moments in W2S voicing decision, expanding our previous use of centroid [17]. The spectral centroid and spread are defined respectively by equations 1 and 2 [24].

\[
\mu_1 = \frac{\sum_k S_k f_k}{\sum_k S_k}
\]

\[
\mu_2 = \sqrt{\frac{\sum_k S_k (f_k - \mu_1)^2}{\sum_k S_k}}
\]

\( k \) is the index of the spectral bin, \( S_k \) is the magnitude spectrum value of the signal frame at bin \( k \), and \( f_k \) is the frequency in Hz corresponding to bin \( k \). A sliding Hanning window of 23ms with a hop size of 6ms (appropriate values for real-time processing) was chosen. Only the phonemic steady parts were considered, since co-articulation can be managed by an interpolation between voicing decisions. Therefore, following a phonetic segmentation of the audio signal (more details in section 2.3), only the 50% central frames of each phoneme were kept for all analyses and classifications.

2.2. Phonemes of interest

This study focuses on French phonemes. Among those, plosives are a particular category that can be segmented into occlusion and release parts. In normal speech, the main difference between voiced plosives and their unvoiced counterpart resides in the occlusion part that may either carry voicing or remain silent. However in whispered speech, plosives are always voiceless and thus can hardly be detected in real-time, i.e. in a causal setup. For this reason, they will not be considered in this study. The phoneme /r/ will not be considered neither since its voicing status may vary in French. Based on these considerations, we will focus on the following phonemes of the French language: vowels, semi-vowels, fricatives, and phonemes /l/, /m/, and /n/. Among the considered phonemes, only fricatives /l/, /s/, and /S/ (using the SAMPA phonetic notation [25]) are unvoiced. However, in whispered speech, voiced and voiceless fricatives share similar characteristics, which make them very difficult to discriminate, both in terms of perception and signal processing [2, 16, 23]. Three types of fricatives with different places of articulation can be identified in French: dental fricatives /s/ and /z/, palato-alveolar ones /S/ and /Z/, and labio-dental ones /l/ and /f/.

Fig. 1 plots the distribution of signal frames from the different phonemes in a 2D map defined by spectral centroid and spread for a French male speaker reading a phonetically-balanced text. It shows that centroid alone is not sufficient to discriminate palato-alveolar fricatives from labio-dental ones, and palato-alveolar fricatives from some vowels like /l/. However, using the spectral centroid and spread together allows to better discriminate the different types of fricatives from each other and from other voiced phonemes. Preliminary analysis confirmed that this behaviour is consistent across speakers, which will be further evaluated in section 4.

2.3. Phonemes Classification

A k-nearest neighbours algorithm (KNN) is chosen to label a signal frame as one of the four pre-defined classes \{[/IZ/\, /IN/] ; [/l/\, /s/] ; [/v/\, /f/] ; [vowels, semi-vowels/l]/[m]/[n]/[a]]\}. In the following, the last class is called all voiced. Particular advantages of the KNN algorithm is that it does not require to make assumptions about data distribution, and can be used with a small training set of annotated signal frames. Based on this training set, a new unknown signal frame will then be assigned the most represented class among its K nearest neighbours in the training data. Note that for KNN classification, centroid and spread values are normalised in the range \([0,1]\) based on the minimum and maximum values of the training set, so that both dimensions have an equal weight when computing the Euclidean distance between data points to find the nearest neighbours. The background colours in figure 1 show how each point of the [centroid x spread] space would be classified using the displayed distribution as the training set.

2.4. Intra-class centroid-based voicing decision

Based on the previously-established classification, each phoneme class can then be treated independently. For signal frames labelled as all voiced, voicing is always applied. For the others, we use a dedicated centroid threshold as criteria for each fricative class, computed as the mean centroid value of the class distribution. If the centroid is above the threshold, the frame will be labelled as unvoiced, and if it is below, it will be
3. Corpus and testing conditions

Due to the lack of existing publicly-available French database of annotated whispered speech, we recorded and annotated a new dataset to evaluate the proposed approach.

3.1. Test data

For assessing the intelligibility of pathological voices, [23] conceived a text in French that covers many phonetic, phonological and linguistic criteria, among which: a full phonetic coverage with a phonetic balance similar to that of French language; multiple apparition of the cardinal vowels /a/, /i/, and /u/; appearance of the unvoiced fricative /s/ followed by each cardinal vowel; rapid occurrence of multiple unvoiced fricatives. This text can be found in [23, p.114], along with the full list of covered criteria. We assume that evaluating the performance of our method on this text should give a good overview of what may be expected in real usage condition, and thus used this text as test data in all of the following evaluations. The five first sentences of the text were removed from the test data, since they are used as one of the training conditions, as detailed below.

3.2. Training conditions

The creation of the training data raises three questions: 1) the size of the database and phoneme coverage; 2) the need to train the system specifically for each user based on individual recordings, or if a single generic system trained on a multi-speaker database would be sufficient for any new unseen user; 3) the quality of training data annotation (automatic vs. manual). All three considerations are trade-offs between classification accuracy, and recording/annotation time. In particular, if it were demonstrated that the proposed method performances would benefit from using personalised training data and manual annotation, then a calibration procedure requiring some recordings from the new speaker followed by data annotation should be performed for each new user. This would not be desirable in terms of workload for both user and experimenter. Therefore we evaluated here the influence of the phonemic content: single-speaker vs. multi-speaker training; and automatic vs. manual annotation, on classification accuracy.

3.2.1. Phonemic content

Four training datasets were created as described below and whose phonemic contents are summarised in Table 1:

- “ph”: Steady phonemes, covering each vowel and fricative, sustained for about 2s (19 items).
- “VICV1”: Sequences of 3 phonemes, where V1 covers all vowels, and C covers all fricatives (78 items).
- “VICV2”: Sequences of 3 phonemes, where V1 and V2 is a pair of different cardinal vowels (/a/, /i/ and /u/) and C is a fricative (36 items).
- “text”: The first 5 sentences of the text mentioned in section 3.1, which have been conceived such that they already cover many phonetic criteria, including a good phonetic balance and full phonetic coverage (63 words).

All the test and training data were recorded by 10 different speakers (5 female and 5 male). Recordings were done in an anechoic chamber using a high-quality DPA4088 headset microphone along with a Complete audio 6 audio interface, with a sampling rate of 44.1kHz and a bit depth of 32 bits.

3.2.2. Speaker and annotation

In order to evaluate the need for an individual system calibration, we compared a single-speaker setup where only recordings of the subject being evaluated are used as a train set, to a multi-speaker set-up where the model is trained on recordings from all speakers except the one being evaluated.

We used the Montreal Forced Aligner [26] for data annotation, with the “french_prosodylab” model available for French language to perform an automatic phonetic segmentation of the recordings. All annotations of the test set were manually corrected using Praat [27] to ensure a robust ground truth to be used for evaluation. To assess how potential errors of the automatic annotation may affect classification accuracy, single-speaker setups were tested both with and without manual annotation. Since a multi-speaker training is done once and for all, we assume that manual annotation is worth doing in this case. To sum up, the three speaker and annotation conditions called training type in the following are: “ind_auto” for single-speaker automatic annotation; “ind_man” for single-speaker manual annotation; “generic” for multi-speaker manual annotation.

3.3. Methods

We compared the KNN-based method to a baseline in all the conditions described above. Based on preliminary tests with KNN, we set the value of K to 51 for all setups and datasets, which gave in average the best results for classification accuracy. A similar approach to that from [17] was used as a baseline, with a single global threshold on the raw spectral centroid of a signal frame to determine voicing. For this method, we set the threshold as the one that maximises the global voicing decision accuracy on the whole training set, as illustrated in Fig. 1.

Overall, the evaluation was performed on 2 methods × 4 training sets × 3 training types. In the following, voicing accuracy is calculated as the ratio of correct voicing decisions computed on each individual test speaker for each method, training set and training type. Results are reported for the 10 test subjects, and multiple comparisons between conditions are performed with a logistic regression.

4. Results

4.1. Comparison of training conditions

Fig. 2 shows the global voicing accuracy obtained on the whole test set, grouped by training set and training type, methods combined. Significance between conditions are indicated by stars on the Figure. Except for the “ph” condition for which individual training is better, the obtained results show that there is little benefit (< 0.2% improvement) in using a user-specific calibration procedure, compared to a generic calibration from several unseen speakers, which greatly simplifies the usage of the
4.2. Comparison of methods

Although both KNN and baseline perform similarly in terms of overall voicing accuracy, some more in-depth analysis shows interesting differences for specific phonemes. Figure 4 (a) shows the confusion matrix of the phonemes classification obtained with our approach trained in the “generic” condition on “text” data. Such representation helps to better understand how classification errors impact voicing decision accuracy, since only misclassified phonemes contained in the red areas imply voicing decision errors. For instance, the classification of /v/ in the all_voiced group does not lead to a voicing decision error. However intra-class decision errors (e.g. /S/ labelled as /Z/) do lead to a wrong voicing decision. Figures 4 (b) and (c) compare the voicing decision error on each phoneme for the proposed and baseline methods, respectively. While both methods present similar results for /v/, /f/, /z/ and /s/, the most striking difference is on the phonemes /S/ and /Z/. They are given an accuracy of 0 and 1 by the baseline, respectively, while errors are more evenly distributed with our method. Fig. 1 illustrates well this result: the position of the baseline threshold classifies all the phoneme class /[S]/ as voiced. In a real-time speaker adaptation paradigm, this set up would not allow the speaker to increase the /S/ centre of gravity above the threshold. Inversely, the class-specific threshold allows much more flexibility to adjust the fricative centre of gravity to enforce the voicing decision when necessary.

5. Conclusion

A new method for automatic voicing decision in W2S conversion was proposed. It first classifies signal frames into three types of fricatives or voiced phonemes based on spectral centroid and spread values. It then further discriminates voiced from unvoiced fricatives based on intra-class centroid thresholds. We first showed that individual system calibration can be avoided by training the algorithm on a pre-annotated multi-speaker database of read text. Second, both proposed method and baseline displayed a voicing accuracy higher than 91% when trained on 9 unseen speakers uttering a phonetically-balanced 63-word text. Third, compared to the baseline, our method allows to reduce systematic voicing errors for some phonemes, opening the path to a more suitable control space for voicing decision. We believe that speakers can learn to adapt to the class-dependant spectral centroid of fricatives in a real-time setting and compensate for the system voicing errors. This hypothesis will be investigated in future work.

6. Acknowledgement

This work has supported by Agence Nationale de la Recherche (ANR-19-CE28-0018). The authors wish to thank Silvain Gerber for his help with the statistical analysis of the results.
7. References


