SPEECH COMPANIONS: EVALUATING THE EFFECTS OF MUSICALLY MODULATED AUDITORY FEEDBACK ON THE VOICE

Rébecca Kleinberger, George Stefanakis, Sebastian Franjou

MIT Media Lab
75 Amherst Street, MA, Cambridge
rebklein@media.mit.edu, stefanag@mit.edu, sfranjou@mit.edu

ABSTRACT

Changing the way one hears one's own voice, for instance by adding delay or shifting the pitch in real-time, can alter vocal qualities such as speed, pitch contour, or articulation. We created new types of auditory feedback called Speech Companions that generate live musical accompaniment to the spoken voice. Our system generates harmonized chorus effects layered on top of the speaker's voice that change chord at each pseudo-beat detected in the spoken voice. The harmonization variations follow predetermined chord progressions. For the purpose of this study we generated two versions: one following a major chord progression and the other one following a minor chord progression. We conducted an evaluation of the effects of the feedback on speakers and we present initial findings assessing how different musical modulations might potentially affect the emotions and mental state of the speaker as well as semantic content of speech, and musical vocal parameters.

1. INTRODUCTION

This work seeks to assess how different musical feedback modulations might affect the general mental state of the speaker, semantic content of speech, emotions in vocal tonalities and vocal parameters of musicality. Modulated Auditory Feedback uses digital signal processing to transform the way someone hears their own voice. Modulated Auditory Feedback has documented effects on how someone speaks in terms of speed, articulation, and fluency. For example adding a short delay to the voice can lead to prolongation of vowels, repetition of consonants, increased intensity of utterance, and other articulatory changes [1, 2]. A short delay (20-150ms) can help people who stutter become more fluent [3] but a longer delay (higher than 200ms) can lead to jammed speech [4].

In recent years, the research community has investigated the possible effects of altering vocal auditory feedback for regulation of emotions [5, 6]. In these studies, modulated feedback is used covertly to make the voice sound more calm, sad, happy or fearful by manipulating formants, overall pitch, and by adding filters. The researchers then established the effects on the subjects measured through self-reported emotions and levels of stress. Our approach consists of producing aesthetic musical manipulation of the voice instead of covert intonation and testing the effects on the speakers fine-tuned ability to shape their voice and speech. Musically Modulated Auditory Feedback is a new approach that creates real-time musical transformations of the voice, for instance by generating guitar chords accompanying the rhythm and pitch variation of the voice, or by creating several pitch shifted versions of the voice and layering them in real-time to create a choir-like harmonization of the voice. We conducted a study to assess whether specific Musically Modulated Auditory Feedbacks can induce particular effects and modulate emotional content from the voice, in addition to affecting vocal parameters. Our objectives are twofold: first, we are interested in studying the potential regulatory effects of music when woven into voice. Second, we wish to bring more awareness to the intrinsic musicality present in everyday speech and explore possible research applications based on perceiving the spoken voice as an inherent musical signal. These applications range from infant-directed speech and language acquisition to speech pathology and aphasia re-education. Such research could also show useful for music composer or could lead to new tools for phonologists to characterise human speech. We present the background supporting our inquiries in terms of neurology, research on music and emotion, and self-perception theory. Then we present the study design and detail the data analysis and results.

2. BACKGROUND

2.1. Musicality of everyday speech

Speech is one of humanity's richest and most ubiquitous forms of communication. Its richness lies in the combination of linguistic and nonlinguistic information. Musicality is a crucial nonlinguistic component of speech, incorporating the tempo and rhythm of the speaker along with the pitch variation and unique texture of vocal sound. In casual everyday speech, individuals possess a unique musicality, rhythm and melodic style. In 1954, urban folklorist and sound archivist Tony Schwartz proposed the idea that "there is music in everyday speech, and often a kind of poetry in the way people talk" [7]. In our work, we aim to increase awareness of the beauty and diversity of musicality in our everyday experience of voices. Vocal, non-verbal behaviors such as prosody, tone, loudness, breathiness, accent, pitch envelope, and tempo are all parameters that are most often unconsciously controlled when speaking, but they implicitly convey a great deal of information. For instance, pitch intervals can reveal changes in mood [8] or hormone levels [9], tempo information can be a marker of depression [10]. Prosody and especially pitch accentuation can also be used to modify semantic content [11].

Our system creates different types of musical layers on top of the spoken voice by extracting existing musicality from speech and aims to bring more awareness on this intrinsic musicality.

https://doi.org/10.21785/icad2019.035

2.2. Music and Emotion

The influence of music on emotion is not a novel concept. As early as 350 B.C., Aristotle characterized different musical modes by the emotions they evoked [12], and throughout the classical age of music, the "feel" of a piece was often married to more objective qualities like tempo and chord. In terms of valence, minor keyed pieces and melodies are traditionally associated to sad, nostalgic or morose atmospheres, including Chopin's Funeral March and Mozart's Requiem. On the other hand, major keyed-pieces are classically associated with joyful, strong and uplifting athmospheres, including Mendelssohn's Wedding March and Rossini's William Tell Overture. Whether some innate qualities of the major and minor tonalities informed theory and popular opinion, or vice versa, is a philosophical inquiry which is not to be dwelled upon, but the popular social perception of the major and minor chords, for hundreds of years in the western tradition, has been as that the former is classically joyful, while the latter is often considered sorrowful [13]. Of course there are exceptions; many pieces of music exist which do not follow this categorization. Furthermore, the perceptions of individual pieces can vary widely from person to person. The famous paper published by Hevner et al. in 1935 elucidates the various affective qualities of the major and minor musical modes [14]. The author claims that major is dynamic, more natural and fundamental than minor, and "expresses varying degrees of joy and excitement." She goes on to assert that "[the major] sounds bright, clear, sweet, hopeful, strong, and happy," while the minor "expresses gloom, despair, sorrow, [and] grief," and is "mournful, dark, [and] depressing." Many theories and studies have supported this notion of musical modes having intrinsic emotional connotations implicit within them, and several support the idea that music can indeed evoke strong emotional responses in listeners [15, 16].

Although the findings that minor chords have a negative valence effect have been presented in many prior work on music emotion, as of the time of this writing, we haven't found any prior work assessing effects of the use of minor vs major keys when interactively woven into spoken voice. In this work, we are proposing a step toward assessing unconscious effects of auditory musical transformation of speech.

2.3. Self-Perception Theory

In his self-perception theory, Daryl Bem [17] postulates that individuals come to know their own attitudes, emotions, and other internal states partially by inferring them from observation of their own overt behaviors. He argues that internal cues are "weak, ambiguous or uninterruptible", and that we often have to rely on external cues to understand our own behaviors the same way an outside observer would.

This theory suggests that it is partly by monitoring the way we overtly express our emotions that we infer our emotional state and attitudes. Multiples studies support this theory, by showing that forcing the outside symptoms of an emotion can reinforce said emotion in the subject [18]. Similar results have been obtained for vocal expression of emotion: subjects asked to imitate vocal patterns associated with specific emotions (eg. laughter) reported their emotions being affected accordingly [19]. The previously mentioned studies involve active cooperation from the subject, but further studies have found similar effects in cases where patients didn't have to consciously adjust their behavior, or weren't even aware of anything being modified. Subjects who heard their voices processed in real-time to make it sound as if they were happy, sad, or afraid experienced changes in tension and self-reported positivity usually associated with the experience of the corresponding emotion. This suggests an influence of the perception of the subjects own voices alone on their emotions despite them not even noticing any modification of their voices [6]. Similarly, participants whose voices were modified to sound calmer and fed back to them in real-time during relationship conflicts reported feeling less anxious than those having unmodified feedback [5]. These studies suggest that emotions can be regulated by feeding back modified version of a speaker's voice in real time even if the modification is not consciously detected.

In our work, we explore this field by modifying the subjects' fed back voices to match purely musical expressive features. Links between prosodic and musical emotional features have been suggested, such as the use of the interval of a minor third for affects of negative valence for both speech and music [20]. By mapping the fed-back voice to musical attributes considered happy or sad we hypothesize similar emotional responses to those induced by previous non-musical manipulation.

2.4. Neural Basis

A large body of work conducted on neural control of speech has been accumulated in Frank Guenther's book of the same name. A key idea presented in the book is that of neural auditory feedback control, which is operated by means of a feedback/feedforward mechanism. In this scheme, it is suggested that fluent speech is dependent on fluid, logical, sensory feedback streaming back to the speaker. It is for this reason, Guenther asserts, that delayed auditory feedback results in a range of dysfluent behavior, up to and including complete cessation of speech [21]. The importance of auditory feedback in speech production has been further proven by studies on the effect of modified real-time and delayed feedback on speech and sustained vowel sounds. It was found that modification of the fundamental frequency (F0) of the feedback voice produces a compensatory opposing shift in the pitch of the resultant sound for both sustained vowels and speech [22, 23] due to brain overcompensation. Formant shifts in feedback have also been found to produce compensatory changes in the spectral characteristics of the voice [24], even when participants were consciously informed of the modifications and instructed not to compensate [25]. Thus it appears that auditory feedback plays a crucial role in speech production, to the point where it sometimes cannot be ignored even if the speaker is consciously trying to combat its effects. The neurological basis of our study is to interfere with the encephalic speech-feedback mechanism by overlaying the stream of one's own raw voice, using musical modulations. The goal is to monitor the alterations in the resulting feed-forward mechanism of new speech being produced. We also seek to analyze the semantic nature of speech produced when the backward-fed vocal audio is substantially altered in either major or minor chord progressions.

2.5. Measure of Musical Parameters in Speech

It can be difficult to assess and characterise the musicality of speech. The question is so polemic that quite often, researchers assess the level of musicality by asking experts with extensive music training to subjectively rate vocal sound samples. In Music, Language and the Brain [26], Aniruddh D. Patel distinguishes musical and linguistic sound systems in the way they carry pitch, timbre, rhythm and melody. One way to assess pitch is through the

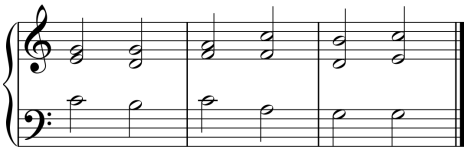
analysis of the mean pitch (Pm) of a vocal sample. Pm provides information about the fundamental frequency (F0) of a subjects voice. Males with lower voice will have smaller F0 thus lower Pm. Level of melodicity can be very roughly accessed through the measurement of the pitch standard deviation (Psd) of a vocal sample. Psd gives cues about the pitch envelope in speech: the lower the Psd in a given phrase, the more monotone and concentrated around the main pitch, the voice will be. Contrary to a lot of musical systems and instruments that present a fixed timbre, speech is also fundamentally a system of organised timbral contrast, as timbral variation in vocal sound is the basis of phoneme production. In addition, on account of the shape of formants, subtle vocal timbral variation is what allows us to distinguish the voice of different speakers. Timbre in speech can be measured with different parameters such as jitter, breathiness, or harmonic-to-noise ratio (HNR expressed in dB). HNR is a more global way to see timbre as it indicates the energy concentration of the sound around the main pitch. HNR represents the degree of acoustic periodicity. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise. And a HNR of 20dB means that 99% of the energy of the signal is in the periodic part. Singing voices have higher HNR than spoken voices [27]. Pm, Psd and HNR are used in this study as measurement of variation of musical parameters of speech. In this work, we did not intend to measure rhythm in speech partially as it is used in the generation of the pseudo beats in speech companions.

3. SPEECH COMPANIONS

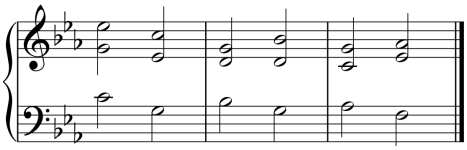
We created new types of auditory feedback called Speech Companions that generate live musical accompaniment to the spoken voice. The Speech Companions used for this study are based on a type of active harmonizer. An harmonizer is a pitch shifter that combines the shifted pitch version with the original sound to create a two or more notes harmony. Our system combines the original vocal signal with two extra layers creating a musical chord. A constant harmony chord being played in a sustained manner can create a very dull effect. In live or studio music production, harmonizers are often controlled manually by a keyboard that changes chords to make it more reactive. For our study, we wanted the feedback to react to the inherent rhythm of speech. Our system triggers a new chord, from a predetermined set, at each pseudo-beat of speech.

Pseudo-beats are triggered at near-regular intervals determined by minimum delay and natural attacks in the voice. Sound attack corresponds to onset or peak in the intensity of the sound signal. After each chord change, the system counts down the chosen delay in milliseconds and then waits for the next speech onset to generate the next pseudo-beat controlling the next chord change. When chords are changed at a regular interval, the feedback seems very static and creates a ticking clock effect that can feel stressful and alter the natural speech rhythm. By using the pseudo-beat method, we ease the chord variation into the organic speech tempo to respect the built-in musicality of speech. The system was implemented using Max MSP 8 for pseudo-beats detection and with MHarmonizerMB for Reaper64 to create the harmonization.

The system randomly draws a chord to harmonize from a predetermined chord progression - either major or minor. The chord progressions were chosen to unambiguously convey the key and mode regardless of which order the chords were played in, as they were to be fed to the subjects in random order. The key of C was chosen, and the chords are in the modes of C ionian (major) and aeolian (natural minor) (see Figures 1.a and 1.b). Although commonly used by classical composers, the harmonic minor was avoided as the augmented second interval can sound jarring or exotic to western listeners. This interval is usually avoided by following proper voice leading rules, but this wasn't possible due to the random order of the chords. The aeolian or natural minor mode, commonly found in popular music, was chosen instead to bypass this problem. The chords are voiced in the mid-range so that the harmonized feedback would not sound too distant in pitch from the normal voices of most subjects. The range and spread of the chords were kept comparable (see Figure 1). The major chords are triads, while the minor chords are sometimes enriched to convey more tension and sadness.



(a) Major chord progression



(b) Minor chord progression

Figure 1: Chord progression used for for the major (a) and minor (b) mode of our study

Figure 2 illustrates the use of the pseudo-beat to trigger changes in the MIDI track that always last longer than a minimum delay and are ultimately triggered by speech attacks from the raw voice. The result generates harmony changes in the processed voice (middle track) that exhibit different spectrum peaks than the raw voice. In this case each chord lasts a minimum of 3000ms but can extend longer if no attack is detected. The volume was kept the same for all participants and was loud enough to mask the actual voice. We hypothesise that such feedback might affect the valence of the speaker as well as the musicality of their speech.

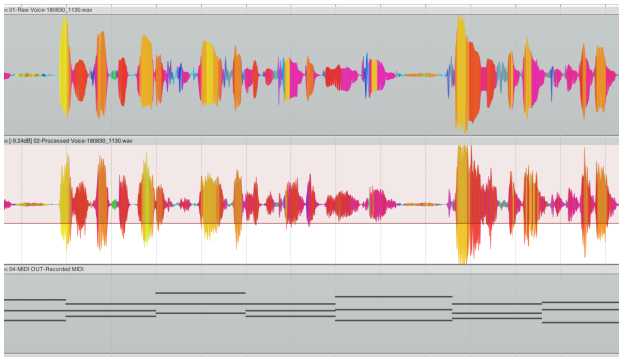


Figure 2: Illustration of the Speech Companion in use: attacks in the raw voice (top track) trigger the midi chords (bottom tracks) that control harmony changes in the processed voice (middle track)

4. STUDY DESIGN

4.1. Participants

The institutional review board approved this study, which was registered as COUHES protocol no.1802248976. The sample comprised 20 adults (11 women and 9 men). There were two groups: one group of 10 adults received the "major scale" condition, and the other group of 10 adults received the "minor scale" condition. No compensation was offered to the participants. The study was organized over 5 days, in which we measured respectively 1, 3, 6, 5 and 5 participants. The settings were identical throughout the 5 days in terms of environment, microphone settings, audio loudness, and lighting.

4.2. Study Setup

The study was conducted in a soundproofed room to reduce background noise. We used a Countryman E6 directional ear-set microphone and a Babyface RME Pro audio interface connected to a computer to record the voice and a pair of Bose SoundSport earphones to provide audio feedback. The SoundSport are very open (i.e. let outside sound in) which allowed the interactions between the subject and the researcher to remain natural. The researcher giving the instructions also wore a pair of SoundSport to monitor the quality of the feedback heard by the subject. The loudness of the feedback was set just loud enough to effectively cover the speakers voice without sounding unnaturally loud.

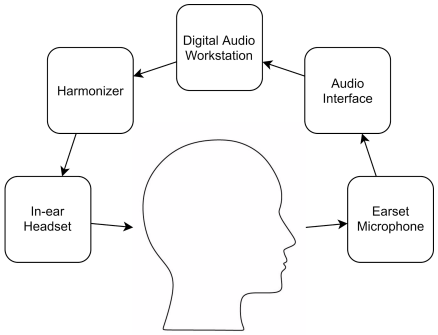


Figure 3: System Flow

4.3. Method

The study was composed of two phases (baseline and musical feedback) each containing the same three tasks (reading, mood assessment and storytelling). Subjects were initially fitted with in-ear headphones and a microphone. During phase 1, subjects did not hear any feedback through the headphones but still had to wear them to get accustomed to it in preparation for phase 2.

- Task 1 is a reading exercise to normalise the subjects mood to neutral at the start of the study. To this end, we use a adapted version of the Velten mood induction process (Velten MIP) method [28]. As we want to induce a neutral mood to all participants, we ask them to read a series of 15 trivial and factual statements which carry no emotional load extracted from the 50 sentences used in Velten MIP version used by Isen and Gorgoglione [29]. This reading task aims at initiating the same neutral common ground for each subject.

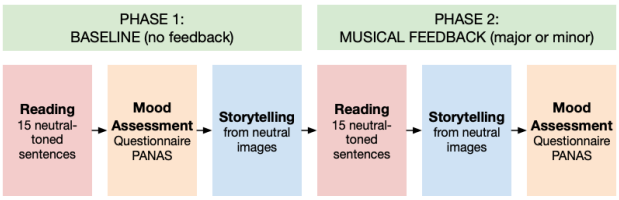


Figure 4: Order of the Study

- Task 2 consisted of filling out a short mood questionnaire to measure self-reported affect. We used the Positive and Negative Affect Schedule (PANAS) methodology [30]. The PANAS was chosen for its robustness, replicatibility, and widespread use, to allow for easy comparison with other works. To limit demand effects, it was issued in digital form where only one question was visible at a time. This prevented the subject from seeing the whole questionnaire and influencing to global results by correlating their answers to several questions.
- In task 3, subjects were shown four images from the IAPS image database and asked to construct a narrative loosely based on the images. IAPS is a database of images for experimental investigations of affect [31]. Each of the chosen images were in the valence range 4.5-5.5, signifying emotional neutrality. A scale score of 1 on the valence portion of the IAPS image scale means unhappy, while 9 means happy. Images with neutral-evoked emotions could go either towards the more joyous, or the more depressed, semantically and tonally.

For the entirety of phase 1, no audio feedback was played through the headphones, though audio from the microphone was being actively recorded. This initial neutral portion of the protocol was used as a baseline to evaluate the effects of the musical modes.

The study then entered phase 2, where the subject had to repeat the same three tasks while hearing the Musicalised Modulated Feedback. For each subject, either the minor or major chord harmonizer was tested and each subject listened to their voice modulated at a volume sufficiently high so as to mask their own voice.

- For this phase, task 1 was composed of 15 new neutral sentences to read.
- for task 2, the subjects were given four different images from the IAPS image database, from which to generate a new story.
- and for task 3, the subject was asked to fill a new randomised PANAS to fill out.

In phase two tasks 2 and 3 are switched compared to phase one as we wanted to give more time to the subject to get used to the feedback before measuring their self-reported mood in order to get a better sense of the change of mood induced by the study.

The musical modulations were then turned off and we asked the subjects their best guess about the purpose of the study to determine if they were aware that their mood and tone were being investigated. Indeed, past research has shown that results of studies on affect might be skewed or unintentionally affected if subjects are aware that their mood is being monitored [32]. At the end of the experiment we then verified that all the participants had remained unaware that the study was about affect and we informed them of the actual objective through a short debriefing session and asked not to divulge it to other potential participants.

5. DATA ANALYSIS

The collected data were processed into three categories: the self reported PANAS result were processed into numerical data. The stories generated (two per subjects) were analysed in two different ways: as text to assess semantic content, and as speech audio sample to assess changes in vocal affect and musicality.

5.1. Self-Reported Affect

The PANAS questionnaire was completed by the subject twice: once as part of the baseline evaluation, and once at the end of the musical-feedback task. The questionnaire gives us scores for positive affect (PA) and negative affect (NA), that are subtracted to obtain a valence score V normalize between -1 and 1. To limit the variations due to differences in initial mood between subjects, we analyzed the variation in valence induced by the experience by subtracting the valence prior and post study. This allowed us to only take into account mood changes from baseline induced during the study. These change in valence was then compared between the minor scale group and the major scale group.

5.2. Semantic Content

To analyze the semantic content of the speech, the audio recordings of the constructed narrative based on the pictures from IAPS in tasks 1.3 and 2.3 were all transcribed to text using Dragon NaturallySpeaking [33], and the text outputs were then reviewed manually and corrected to assure accurate transcription of speech. These text transcriptions were processed using the SentiWordNet database which scores words based on their positivity and negativity [34]. For each subject, we compared the difference in average positive, negative, and total scores from the SentiWordNet analysis between the baseline story and the story invented by the subject while hearing musical feedback.

5.3. Emotion Analysis from Vocal Intonations

Emotional vocal qualities were analyzed using the speech emotion recognition software OpenVokaturi [35]. OpenVokaturi is a Software Development Kit developed by Vokaturi to provide analysis of the basic emotions from speaker's vocal intonations. It is worth noting that the SDK is presented as having an accuracy on classification of only 66.5 % which highly limits the validity of the results [36]. Vokaturi provides percent likelihoods for neutrality, happiness, sadness, anger, and fear. Each speech audio sample was analyzed using the OpenVokaturi pretrained model. Scores for positive and negative affect were constructed by way of a weighted sum (Positive Affect = Happiness; Negative Affect = (Anger + Fear = Sadness) / 3), in a similar fashion to the PANAS's way of summing different positive and negative reported emotions to construct positive and negative affect [30]. We then took the differences between the scores for speech segments produced under the musically modified feedback and those produced under normal feedback conditions. To mitigate the effects due to subject particularities, we considered the relative change in affect between the baseline phase and the musical feedback phase rather than absolute affects

5.4. Vocal Parameters

We used Praat [37] for the analysis of vocal and musical parameters of speech. For the speech samples of the narative generated by subjects in phase one and two, we extract mean pitch (Pm), pitch standard deviation (Psd) and harmonic-to-noise ratio (HNR) of the voiced sections of speech. A vocal sound is said to be "voiced" when it originates from the vocal chord and not only from air leaving the lips (e.g. all vowels and diphthongs are voiced, consonants can be either voiced or unvoiced). The analysis parameters were set in Praat as followed: pitch was computer by autocorrelation between 44 and 400Hz with an octave jump cost of 3.5 on voice sections defined with a silence threshold of 0.05 and a voicing threshold of 0.25 and a voice/unvoice cost of 0.15. Detected pitch were also visually validated by researchers. For this section, we hypothesise that whatever the mode (major or minor), speech from phase 2 might have different Pm, Psd and HNR than speech from phase 1.

6. RESULTS

We report findings on data comparing changes in valence between the major scale and the minor scale group as well as changes of vocal parameters (Pm, Psd and HNR) induced for both group by the experience. All t-tests were preceded by an F-test to determine whether the samples should be assumed to have equal or unequal variance and the relevant paired t-test was then run accordingly. The significance level of all tests was set to p = 0.05

6.1. Results from Self-Reported Affect

We hypothesized that the minor mode would induce a more negative mood, and that the major mode would induce a more positive mood in the subjects. This was evaluated in three different ways. The first was self reported mood by means of a digital version of the PANAS questionnaire. We observed trends concurring with our hypothesis as the average in valence change was higher for the major scale group (3.3%) than for the minor scale group(1.2%) However a two-tailed T-test didn't show statistical significance. It is interesting to note that both groups general mood seemed to slightly increase after the study (with major mode increasing more than minor mode) which might be due to the surprise and novelty effect

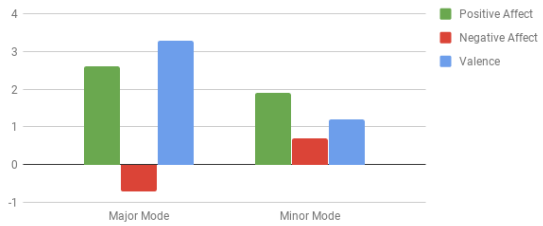


Figure 5: Difference in self-reported positive and negative affect

6.2. Results from Semantic Content

The semantic score analyses conducted on major and minor chord progressions centered around positive, negative, and valence word scores, which give holistic, normalized, numerical attributes of the degree to which the words spoken by a subject leaned more towards positive or negative speech. The valence score was calculated as the sum of the positive and negative scores. We used the Natural Language Toolkit (NLTK) [38] to obtain these scores, and the text was obtained from subjects image narratives, from phases 1 and 2.

We computed the differences in semantic scores from phase 1 to 2 of the study and then compared these across major and minor modes. We used a two-tailed t-test on the valence results as well as on the positive and negative results, and while our results didn't show statistical significance, they still present the expecting trends. We specifically observed that subjects from the minor group had a negative score difference (difference in holistic evaluation of negative words from phase 1 to phase 2), on average almost 6 times higher than those with the major mode; one-tail two-Sample t(18) = -1.0, p = 0.33 >0.05. Still, we cannot reject the null hypothesis with respect to semantic results.

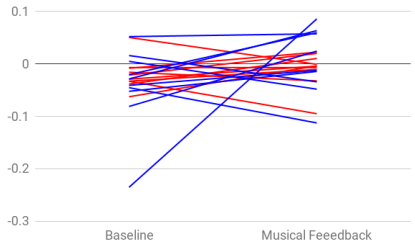


Figure 6: Evolution of semantic score valence (normalised between -1 and 1) between baseline and musical feedback for all participants the blue lines represent the subjects from the minor group and red lines represent subjects form the major group

6.3. Results from Emotion Analysis from Vocal Intonations

The third portion of analysis was comparison of the major and minor groups in terms of emotions extracted from the voice. To accomplish this, we used the speech emotion recognition software Vokaturi. As in the previous analyses, the speech used was obtained from subjects image narratives, from phases 1 and 2. We grouped the normalized Vokaturi data into three areas: positive affect, negative affect, and valence. In accordance with our hypothesis, the negative affect score was found to be significantly greater for subjects subjected to the minor mode compared to those subjected to the major mode. The two-tailed t-test, t(18) = -2.68, p = 0.015 <0.05, agrees with this finding and thus we can reject the null hypothesis here. We also localized this difference to vocal parameters indicating sadness and anger, which implies significantly that the minor mode heightens these emotions in the speaker.

Furthermore, we found that valence, or the difference between positive and negative affect scores, increased on average by over 5 times more for those who had the major mode versus those who had the minor mode; Two-Sample t(18) = 2.76, p = 0.013 <0.05. This serves to show that those who listened to the major mode feedback were much more vocally positive than negative, as compared to those with the minor mode. Although not significant, observed trends also suggest that the major mode increases happiness and positive affect in speakers. The significance of these results should also take into account the relatively low accuracy of the OpenVokaturi tool.

6.4. Results from Vocal Musical Parameters

When analyzing vocal musical parameters, we hypothesized that regardless of key (major or minor), speech from phase 2 might have different Pm, Psd and HNR than speech from phase 1.

A paired-samples two-tailed t-test was conducted to compare Pm between baseline and musical feedback conditions. There was no significant difference in the Pm between baseline (M=154.6Hz, SD=45.6Hz) and musical feedback (M=156.6Hz, SD=47.2Hz) conditions; t(19)=-0.80, p = 0.430 >0.5. This indicates that fundamental frequencies didn't change much in speakers with or without feedback.

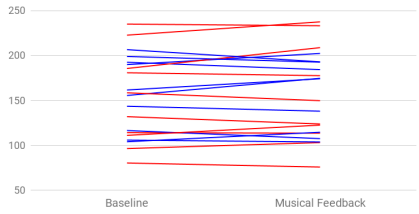
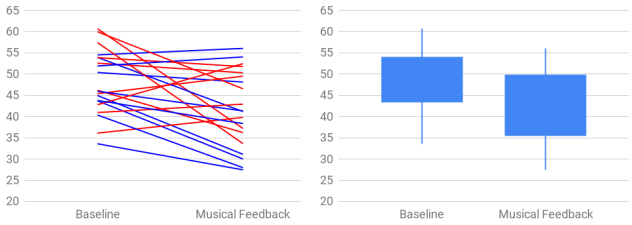


Figure 7: Evolution of mean pitch (in Hz) between baseline and musical feedback for all participants (blue lines for subjects in the minor group and red lines for subjects in the major group

However, significant differences were observed when running a paired-samples two-tailed t-test to compare Psd between baseline (M=47.9Hz, SD=7.5Hz) and musical feedback (M=41.8Hz, SD=9.0Hz) conditions; t(19)=3.024, p = 0.0069 <0.05. This result indicates that speakers became slightly more monotonous and pitch envelopes were less accentuated when hearing musical feedback. We might have expected that musical feedback would make subjects more melodic but instead it seems that as melodic and harmonic matter was added to their speech, they became more conservative in terms of accent, pitch contours and melody in their own produced speech.

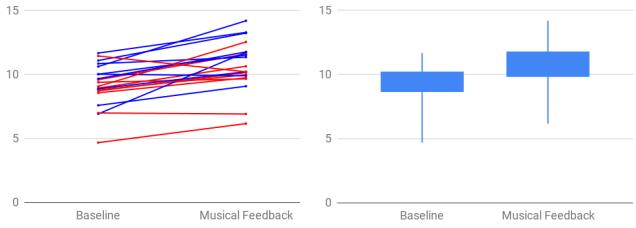


(a) (b)

Figure 8: Pitch standard deviation evolution (in Hz) between the baseline and the musical modes (blue lines for minor group and red lines for major group) (a) and for the entire population (b)

Finally, significant differences were also obtained when running a paired-samples two-tailed t-test to compare HNR between baseline (M=9.2 dB, SD=1.7dB) and musical feedback (M=10.7dB, SD=1.9dB) conditions; t(19)=-5.0, p = 0.000087 <0.05. This indicates that the spoken voice becomes more singing-like with a more precise and accentuated pitch.

Those two results indicate that in terms of timbre, the spoken voice becomes more music-like but in terms of pitch envelope, the speaker becomes more cautious and conservative. This could indicate that the subjects were distracted and further explorations should assess that potential element. This could also indicate that they were paying more attention to listening and integrating their own voice as music rather than language.



(a) (b)

Figure 9: Harmonic-to-Noise ratio evolution (in dB) between the baseline and the musical modes for each participant (with blue lines representing the minor group and red lines representing the major group) (a) and for the entire group (b)

7. DISCUSSION AND FUTURE WORK

When analyzing the data for possible valence and musical effect of musically modulated auditory feedback, we have observed some preliminary results suggesting a trend in the expected direction: self reported valence became more positive for subjects hearing the major mode than for those hearing the minor mode, though not to a statistically significant extent, on account of the small sample size. Analysis of semantic content of speech also didn't show significant results, suggesting that, if present at all, cognitive mood change due to major or minor chords is marginal. However, our study showed significant changes in vocal emotionality and in vocal musicality with a higher harmonic-to-noise ratio and lower pitch standard deviation. This suggests that the feedback makes peoples voice more song-like while reducing their pitch envelope and changes their vocal (but not verbal) emotional content. Additional studies would be necessary to better understand these effects and the factors contributing to them.

This exploratory work presents several limitations both in the context and format of the study. Relatively small sample size and possible order effects are elements that have to be addressed in our future studies. The next stage of the work will also include a different type of baseline where the subject hears their voice amplified at the same loudness without any modulation. Another possible comparison could be with a mode where subjects hear music unrelated with their speech, thought previous studies have indicated that this might create a high level of distraction. Indeed, being subjected to music has been shown to be detrimental to short term memory and cognitive tasks such as reading or processed word writing [39]. In our study, it seemed that the modulated feedback didn't affect to a large extent the ability of the subjects to speak and concentrate. Our subjects seemed sometimes slightly less cognitively and vocally fluent with the feedback but to a lesser degree than one would expect with background music at the same loudness. However, it might still be interesting in the future to assess the level of distraction induced by the system and see how distraction might be reduced when musical stimuli are responsive to user input compared to non-interactive stimuli such as background music.

Further investigations are required to also test factors such as novelty effects, social dynamic related bias, or task induced variability. The musically-altered feedback was quite novel and unusual for many, and some subjects found it, at first, to be amusing or intriguing. Such reactions would tend to indicate an initial boost of positive affect that could then skew the results and mitigate the expected variations, especially in the minor direction. In future explorations, we are also interested in comparing the reaction with Speech Companions when made to adjust to the subject's natural voice range, and see if the system can be made to blend even more with the natural musicality of the voice compared to imposing externally defined musicality onto it. Finally, in this study, the vocal modifications were made obvious, and the subjects were informed of the presence and general characteristics of the modifications. We believe it would be of interest to determine whether a more subtle modification (eg. lower feedback volume) would have comparable, enhanced or reduced effects, similar to the way both pitch shift compensation has been studied for both uninformed [22, 23] and informed [25] subjects. Finally, due to the number of people surveyed and the time frame of the study, we did not include group with no feedback as a control, but in future research we hope to test additional subjects, of which some will not hear any feedback and some will only hear background music while they speak. These extensions would help to buttress the findings of this study.

In terms of real-life applications, we are currently exploring the potential of musically modulated auditory feedback in different contexts, from assistive tools to increase fluency for adults who stutter, to systems for students musicians to better connect with music, to tools for composer who want to explore the musicality of the voice. We also believe that adaptation of such system might be beneficial in new sorts of practices as a possible therapy or relaxation assistant, due to their potential effectiveness in modifying both various voice characteristics and perceived emotion.

8. CONCLUSION

In this study, we created a new type of digital audio manipulation to generate real-time manipulation of the voice through Musically Mediated Auditory Feedback. Classification results significantly indicate that such feedback might alter voice quality and emotion valence detected from voice tonalities. Significant changes in vocal timbre and pitch variation were observed showing the potential to affect speech musicality at a subconscious level.

This early exploration proposed original ways to manipulate the voice in real-time as a way to potentially affect internal mental and physical processes in speakers. By musically altering the way people hear their own voice, we also aim to raise questions about the existing underlying effects of musicality already present in the voice and its reinforcing potential in terms of enhanced emotional regulation, self-awareness, and musicality, in the context of everyday speech.

Speech is one of the richest and most ubiquitous modalities of communication used by human beings. Its richness lies in the combination of linguistic and nonlinguistic information it conveys. Musicality is one of the most crucial nonlinguistic components of speech; it includes the tempo and rhythms of the speaker as well as the pitch variation and unique texture of the vocal sounds. Abstracting musicality from a speech in real time presents several challenges, but explorations in the domain of musically modulated speech and feedback could open doors to to explore real-life situations where the music of speech impacts speakers or listeners such as in the contexts of infant-directed speech, language acquisition, human-animal communication, speech pathology, aphasia re-education, or even music learning and musical composition.

9. REFERENCES

[1] A. J. Yates, "Delayed auditory feedback." *Psychological bulletin*, vol. 60, no. 3, p. 213, 1963.

[2] G. Fairbanks and N. Guttman, "Effects of delayed auditory feedback upon articulation." *Journal of Speech & Hearing Research*, 1958.

[3] J. Kalinowski *et al.*, "Stuttering amelioration at various auditory feedback delays and speech rates," *International Journal of Language & Communication Disorders*, 1996.

[4] G. Fairbanks, "Selective vocal effects of delayed auditory feedback." *Journal of Speech & Hearing Disorders*, 1955.

[5] J. Costa *et al.*, "Regulating feelings during interpersonal conflicts by changing voice self-perception," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 631.

[6] J.-J. Aucouturier *et al.*, "Covert digital manipulation of vocal emotion alter speakers emotional states in a congruent direction," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. 948–953, 2016.

[7] T. Schwartz *et al.*, "Interview with tony schwartz, american hörspielmacher," *Perspectives of New Music*, 1996.

[8] S. K. Blau, "Musicality of speech changes with mood," *Physics Today*, vol. 63, pp. 16–17, 2010. [Online]. Available: http://physicstoday.scitation.org/doi/10.1063/1.4797228

[9] D. R. Feinberg *et al.*, "Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice," *Hormones and behavior*, 2006.

[10] Y. Yang *et al.*, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, 2013.

[11] V. K. R. Sridhar *et al.*, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *Proceedings of Speech Prosody*. International Speech Communication Association Campinos,, Brazil, 2008.

[12] S. Halliwell *et al.*, *Aristotle's poetics*. University of Chicago Press, 1998.

[13] R. G. Crowder, "Perception of the major/minor distinction: I. historical and theoretical foundations." *Psychomusicology: A Journal of Research in Music Cognition*, 1984.

[14] K. Hevner, "The affective character of the major and minor modes in music," *The American Journal of Psychology*, vol. 47, no. 1, pp. 103–118, 1935.

[15] J. A. Sloboda, "Music structure and emotional response: Some empirical findings," *Psychology of music*, 1991.

[16] K. R. Scherer, "Expression of emotion in voice and music," *Journal of voice*, vol. 9, no. 3, pp. 235–248, 1995.

[17] D. J. Bem, *Self Perception Theory*, 1972.

[18] P. M. Niedenthal, "Embodying emotion," *science*, vol. 316, no. 5827, pp. 1002–1005, 2007.

[19] E. Hatfield and C. Hsee, "The impact of vocal feedback on emotional experience and expression," 1995.

[20] M. E. Curtis and J. J. Bharucha, "The minor third communicates sadness in speech, mirroring its use in music." *Emotion*, vol. 10, no. 3, p. 335, 2010.

[21] F. Guenther, *Neural Control of Speech*, MIT Press, Ed., 2016.

[22] T. A. Burnett *et al.*, "Voice f0 responses to manipulations in pitch feedback," *The Journal of the Acoustical Society of America*, 1998.

[23] ——, "Voice f0 responses to pitch-shifted auditory feedback: a preliminary study," *Journal of Voice*, 1997.

[24] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation of speech i: Compensation and adaptation," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 2, 2002.

[25] K. G. Munhall *et al.*, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 384–390, 2009.

[26] A. D. Patel, *Music, language, and the brain*. Oxford university press, 2010.

[27] E. L. Stegemööller *et al.*, "Music training and vocal production of speech and song," *Music Perception: An Interdisciplinary Journal*, vol. 25, no. 5, pp. 419–428, 2008.

[28] E. Velten Jr, "A laboratory task for induction of mood states," *Behaviour research and therapy*, 1968.

[29] A. M. Isen and J. M. Gorgoglione, "Some specific effects of four affect-induction procedures," *Personality and Social Psychology Bulletin*, vol. 9, no. 1, pp. 136–143, 1983.

[30] D. Watson *et al.*, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, 1988.

[31] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, pp. 39–58, 1997.

[32] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: A meta-analysis," *European Journal of social psychology*, vol. 26, no. 4, pp. 557–580, 1996.

[33] "Dragon Naturally Speaking, howpublished = https://www.nuance.com/dragon.html, note = Accessed: 2019-03-26."

[34] S. Baccianella *et al.*, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010.

[35] Vokaturi. Vokaturi. [Online]. Available: https://developers.vokaturi.com/getting-started/overview

[36] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, p. 40, 2012.

[37] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[38] E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[39] P. Salamé and A. Baddeley, "Effects of background music on phonological short-term memory," *The Quarterly Journal of Experimental Psychology Section A*.