Check for updates

SOFTWARE TOOL ARTICLE

REVISED STRAIGHTMORPH: A Voice Morphing Tool for Research in Voice Communication Sciences

[version 2; peer review: 4 approved, 1 approved with reservations]

P Belin[1], H Kawahara[2]

[1]Institut de Neurosciences de la Timone, Aix-Marseille-University, Marseille, Provence-Alpes-Côte d'Azur, UMR7289 CNRS, France
[2]Center for Innovative Research and Liaison, Wakayama University, Wakayama, Wakayama Prefecture, Japan

v2 **First published:** 24 Jul 2024, **4**:154
https://doi.org/10.12688/openreseurope.18055.1
Latest published: 10 Jan 2025, **4**:154
https://doi.org/10.12688/openreseurope.18055.2

Abstract

The purpose of this paper is to make easily available to the scientific community an efficient voice morphing tool called STRAIGHTMORPH and provide a short tutorial on its use with examples. STRAIGHTMORPH consists of a set of Matlab functions allowing the generation of high-quality, parametrically-controlled morphs of an arbitrary number of voice samples. A first step consists in extracting an 'mObject' for each voice sample, with accurate tracking of the fundamental frequency contour and manual definition of Time and Frequency anchors corresponding across samples to be morphed. The second step consists in parametrically combining the mObjects to generate novel synthetic stimuli, such as gender, identity or emotion continua, or random combinations. STRAIGHTMORPH constitutes a simple but efficient and versatile tool to generate high quality, parametrically controlled continua between voices – and beyond.

Keywords

voice; voice conversion; morphing; voice identity; voice affect; animal vocalizations

This article is included in the European Research Council (ERC) gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓ ?

	1	2	3	4	5
version 2 (revision) 10 Jan 2025	✓ view		✓ view	✓ view	? view
version 1 24 Jul 2024	? view	✓ view			

1. **Nadine Lavan**, Queen Mary University of London, London, UK
2. **Yi-Wen Liu**, National Tsing Hua University, Hsinch, Taiwan
3. **Andrey Anikin**, Univ. of Lund, Lund,, Sweden
4. **Stefan Schweinberger**, University of Jena, Jena, Germany
5. **Lee M. Miller**, University of California, Davis, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Horizon 2020 gateway.

This article is included in the Electrical, Electronic and Information Engineering gateway.

Corresponding authors: P Belin (pascal.belin@univ-amu.fr), H Kawahara (kawahara@wakayama-u.ac.jp)

Author roles: Belin P: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Kawahara H**: Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Software, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No.788240). A comparative study of voice perception in primates COVOPRIM, from the French Agence Nationale de la Recherche (ANR-16-CE37-0011-01 PRIMAVOICE; ANR-16-CONV-0002 Institute for Language, Communication and the Brain; ANR-11-LABX-0036 Brain and Language Research Institute), from the British Biotechnology and Biological Sciences Research Council (BBE0039581) and form French Fondation pour la Recherche Médicale (AJE201214).

Copyright: © 2025 Belin P and Kawahara H. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Belin P and Kawahara H. **STRAIGHTMORPH: A Voice Morphing Tool for Research in Voice Communication Sciences [version 2; peer review: 4 approved, 1 approved with reservations]** Open Research Europe 2025, **4**:154 https://doi.org/10.12688/openreseurope.18055.2

First published: 24 Jul 2024, **4**:154 https://doi.org/10.12688/openreseurope.18055.1

REVISED **Amendments from Version 1**

Following Reviewer Lavan's valuable suggestion we have added a paragraph pointing the reader towards additional example uses of STRAIGHTMORPH:

STRAIGHTMORPH provides a relatively simple but highly versatile tool to generate parametrically-controlled, high-quality sound interpolations between voice samples – and beyond –for behavioural, neuroimaging or electrophysiological experiments. Examples of its use include the generation of : voice gender continua (Charest *et al.,* 2013; Pernet & Belin, 2012); continua between different speaker identities including caricatures and "anti-voices" (Latinus & Belin, 2011); emotion continua and caricatures (Bestelmeyer *et al.*, 2014; Bestelmeyer *et al.,* 2017; Giordano *et al.,* 2021; Whiting *et al.*, 2020) averages of increasing number of voices (Bruckert *et al.*, 2010); voice personality continua (Belin *et al.*, 2017); continua between voice and musical instruments (Belizaire *et al.*, 2007); random identities; continua between non-human primate vocalizations, etc. It is our hope that the open access to STRAIGHTMORPH will facilitate research in the emergent field of Voice Communication Sciences.

Any further responses from the reviewers can be found at the end of the article

Why voice morphing?

The ability to parametrically manipulate the physical characteristics of basic sensory stimuli – e.g., the direction of a moving bar on a screen, or the frequency of a pure tone—has played a major role in experimental psychology. But what is easy to perform for simple synthetic stimuli becomes much harder with complex naturalistic stimuli such as faces or voices.

The availability of face morphing software has resulted in important advances in our understanding of the perceptual and cerebral representations of face identity and emotion. Thanks to these tools, researchers can generate synthetic stimuli that preserve essential perceptual aspects of faces and that can be parametrically controlled. For instance, one can generate a continuum of faces between 2 identities, or between one identity and an average face, such that the physical distance between any two consecutive faces is exactly similar at any point of the continuum, allowing a dissociation of physical vs. perceptual differences between stimuli (Koyano *et al.*, 2021; Leopold *et al.*, 2001; Leopold *et al*., 2006; Morris *et al.*, 1996; Rotshtein *et al.*, 2005).

Legacy STRAIGHT

A similar implement thankfully exists in the voice modality: the open-source Matlab package Legacy STRAIGHT has been developed by the second author as a speech vocoder allowing high-quality transformations of the speech signal (Kawahara & Matsui, 2003; Kawahara & Morise, 2024) and its use by an growing number of researchers yields exciting results (Bruckert *et al.*, 2010; Charest *et al.*, 2013; Latinus *et al.*, 2013; Nussbaum *et al.*, 2022; Skuk *et al.*, 2020; von Eiff *et al.*, 2022)}. The vocoder tools decompose input speech signals into sets of vocoder parameters. They are a) the spectral envelope (filter parameter), b) the fundamental frequency (f0), and c) the aperiodicity (time-frequency map of ratio between random and periodic components). To achieve this, STRAIGHT performs an instantaneous pitch-adaptive spectral smoothing based on robust f0 extraction for pitch-synchronous estimation of the spectral profile. STRAIGHT generates a 'smoothed 'spectrogram using a complementary set of windows designed by the extracted f0 and inverse filtering in spline space. The extracted f0 is also applied to separate the periodic vs. aperiodic parts of the vocal signal. The periodic and aperiodic parts of the signal can then be recombined and resynthesized, generally yielding a high-quality synthetic version of the original stimulus. They can also be independently and parametrically manipulated or combined across several samples before resynthesis, yielding novel synthetic stimuli. Although it has been developed primarily as a high-quality speech vocoder, STRAIGHT also works very well for manipulation of other types of vocal information such as identity or emotion, and can work well with vocalizations from other species (Chakladar *et al.*, 2008; Furuyama *et al.*, 2017)

The purpose of the present paper is to make easily available to the scientific community an efficient open-source voice morphing tool called STRAIGHTMORPH (Belin & Kawahara, 2024) and provide a short tutorial on its use with examples. STRAIGHTMORPH consists of Matlab functions from the open-source package Legacy STRAIGHT, plus functions developed in the first author's Voice Neurocognition Laboratory particularly **ExtractmObject** and **VoiceMultiMorph**.

STRAIGHTMORPH provides a relatively simple but highly versatile tool to generate parametrically-controlled, high-quality sound interpolations between voice samples – and beyond –for behavioural, neuroimaging or electrophysiological experiments. Examples of its use include the generation of : voice gender continua (Charest *et al.*, 2013; Pernet & Belin, 2012); continua between different speaker identities including caricatures and "anti-voices" (Latinus & Belin, 2011); emotion continua and caricatures (Bestelmeyer *et al.*, 2014) (Bestelmeyer *et al.,* 2017; Giordano *et al.,* 2021; Whiting *et al.*, 2020); averages of increasing number of voices (Bruckert *et al.,* 2010); voice personality continua (Belin *et al.*, 2017); continua between voice and musical instruments (Belizaire *et al.,* 2007); random identities ; continua between non-human primate vocalizations, etc. It is our hope that the open access to STRAIGHTMORPH will facilitate research in the emergent field of Voice Communication Sciences.

Methods

Operation

STRAIGHTMORPH requires a recent version of Matlab with the Signal Processing toolbox. All files can be accessed for download at https://github.com/pascalbelin/STRAIGHTMORPH. The folder 'scripts' should be added to the Matlab path. The operation flow consists in (i) extracting mObjects for each sound; and (ii) combining several mObjects together to generate novel stimuli.

The TUTORIAL folder contains example sounds and mObjects to be used with the script STRAIGHTMORPH_tutorial.m

Implementation

Step 1 – Extracting mObjects

The function **ExtractmObject.m** calls STRAIGHT scripts to analyze a sound and generate an 'mObject' for this sound. MObjects have been devised by the second author as a convenient way to store all the information concerning each sound to be analyzed, resynthesized and combined with other sounds. They consist of a Matlab structure with 16 different fields. The most important fields are:

(i) mObject.**waveform** containing the stimulus waveform;

(ii) mObject.**F0** containing a vector of length N equal to the duration of the stimulus in milliseconds, with f0 estimates every millisecond;

(iii) mObject.**spectrogram**, containing a 513 (frequencies) × N (millisecond time points) matrix containing the smoothed spectrogram of the stimulus's periodic component;

(iv) mObject.**aperiodicityIndex** containing a matrix of similar size as the spectrogram but containing an time-frequency map of ratio between random and periodic components;

(v) mObject.**anchorTimeLocation** and mObject.**anchorFrequency** containing information on the time and frequencies, respectively, of the Time and Frequency anchors used to combine several mObjects.

ExtractmObject invites **manual user intervention** at two critical steps (also detailed in the script):

(i) **F0 estimation.** Estimation of the time varying fundamental frequency (f0) is a critical step in extraction of the mObject and largely determines the quality of resynthesis and then morphing. The robust f0 extractor in STRAIGHT works well in most cases, but in some cases of e.g. noisy or highly aperiodic sounds the automatic procedure yields unsatisfying f0 trajectories, with e.g. interruptions or shifts to higher harmonics in some portions of the signal. ExtractmObject offers the option to manually limit the frequency range in which f0 candidates are analyzed to try and correct unwanted discontinuities (Figure 1B,C).

Upon calling the function a new window opens displaying a time-frequency graph with different f0 candidates. Note that the mouse should stay on that window to avoid problems when clicking. The f0 trajectory selected by the automatic procedure will be highlighted in green. If the result of this automatic f0 detection looks correct, without unnecessary interruptions, the user can skip the manual adjustment step by pressing the Esc key and move to the next step. If, however, the result of the automatic f0 detection looks unsatisfying, the user clicks on the figure to manually define lower and upper frequency constraints. First the lower boundary is defined by clicking with the left button to add landmarks and with the right button to finish boundary definition. Then the same has to be done for the upper frequency boundary. The F0 detection process will be restarted by STRAIGHT taking into account the manually defined constraints (Figure 1C).

(ii) Definition of time and frequency anchors. A crucial step in face morphing consists in defining landmarks to be put in correspondence across faces, i.e., eyes with eyes and mouth with mouth. Similarly, STRAIGHT makes use of spectro-temporal anchors to constrain the interpolation between different stimuli. Defining accurately time and frequency anchors is of crucial importance for high-quality morphing. Time anchors usually indicate beginning and end of the sound, phoneme temporal boundaries and important formant transitions. Frequency anchors usually indicate formant central frequencies. The number of time axes should be the same for all sounds to be morphed together. The number of frequency anchors can be different for different time axes; yet for each time axis, the number of frequency anchors has to be consistent across all the sounds.

ExtractmObject provides a graphical interface to manually set Time and Frequency anchors (these values can also be edited in the anchorTimeLocation and anchorFrequency fields of the mObject structure). Directly after the f0 extraction stage, a smoothed spectrogram of the analyzed stimulus is displayed full screen. The user has to first define a Time location by clicking with the right mouse button at the appropriate temporal location below the x (time) axis. This will display a vertical time line on which candidate Frequency anchors (local energy maxima) appear at different frequencies as black circles. The user then defines one or more frequency anchors on this time axis by clicking at the chosen frequency with the left button. This process can be repeated as necessary to create new Time locations and selecting new Frequency anchors. The user can press Esc to start the definition process again or Enter to finalize mObject extraction.

Figure 1D shows the smoothed spectrogram of an mObject with its Time and Frequency anchors shown as black circles: the word 'head' spoken by a male speaker (M048eh). Time anchors are positioned at onset and offset of the voiced part of the first syllable, and at plosive onset; Frequency anchors are positioned on the first 3 formants. Anchors can be more complex such those shown for the word 'Hello' spoken a female speaker (Figure 1E).

Once an mObject has been generated, it can be first be inspected using the STRAIGHT function displaymObject. Then it is crucial to check its quality by resynthesizing it and listening to the resynthesized version compared to the original. When mObject extraction works well, the difference between the original and resynthesized versions should be barely perceptible, if at all. A noticeable difference between the two, likely arising from a problem with the f0 extraction step, indicates poor mObject extraction likely to contaminate all morphed stimuli and that should preferably be redone.

Step 2 – Combining mObjects

The function **VoiceMultiMorph** pilots the morphing of an arbitrary number of mObjects. Its general functioning is

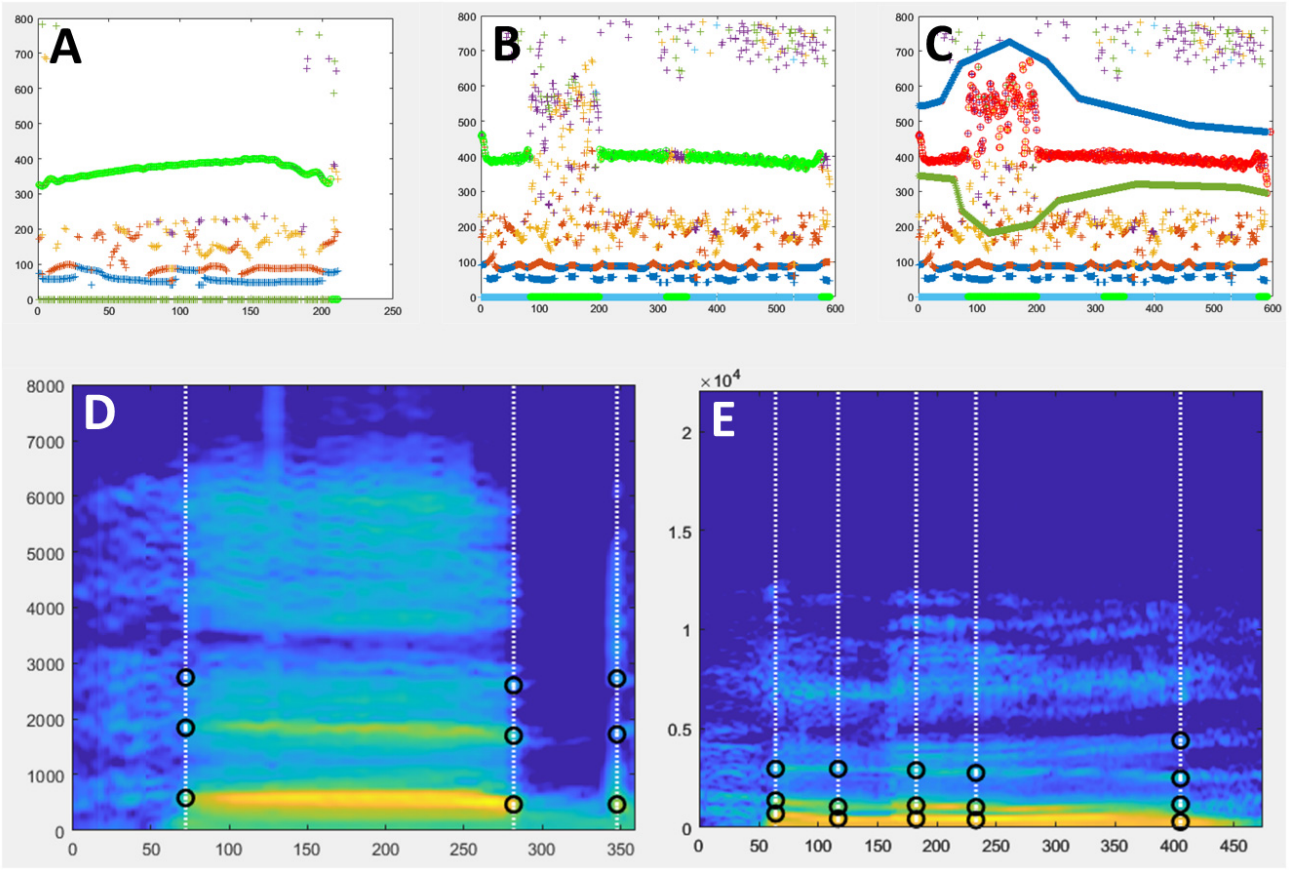

Figure 1. MObject extraction. A. Example of successful automatic f0 extraction using ExtractmObject. The diagram (X axis: time (ms); Y axis: frequency(Hz)) shows identified f0 candidates as coloured circles. F0 candidates selected by the automatic procedure are highlighted in green. **B.** Example of problematic automatic f0 extraction. The green f0 contour is interrupted at several instances which will cause artefacts at resynthesis. This problem can be minimized using manual edition of f0 search range. **C.** Example in B for which f0 search range has been restricted manually to minimize interruptions. The new selected f0 contour is shown in red. **D, E.** Examples of Time and Frequency anchors (black circles) shown on the smoothed spectrograms of two mObjects to be used in examples below. **D**: word 'head' spoken by a male speaker. **E**: word 'Hello' spoken a female speaker.

explained below then illustrated by different examples. All mObjects to be morphed are processed in a symmetrical way, without intermediate morphing stages. The mObjects to be morphed are first loaded in a cell array of structure. A critical step then consists of defining the *morphing rates* ('mRates') for each sound. mRates can be thought of as the coordinates of the sound to be generated in a multidimensional voice space with as many dimensions as original mObjects. They consist of a structure with 5 fields, each of them being a vector with as many elements as mObjects indicating the weight of each mObject for interpolation of that particular field. The 5 fields are: (i) F0: mRates vector for f0; (ii) spectralamplitude : mRates vector for spectral amplitude; (iii) aperiodicity: mRates vector for aperiodicity index; (iv) time: mRates vector for Time anchors; (v) frequency: mRates vector for Frequency anchors. For each of the 5 fields the sum of all mRates must equal one for that field. mRates for a given sound can be different across the 5 fields (cf. 'Parameter-specific morphing' below).

Default settings for interpolation based on mRates are logarithmic for f0, spectralamplitude and Frequency anchors; linear for Time anchors and aperiodicity indices (which are on a dB scale). Calling the function Voicemultimorph will generate a new mObject ('mObjectM') based on the mRates provided for the different mObjects, which can then be examined and resynthesized as any other mObject.

Morphing examples

Morphing 2 mObjects

The examples below details the interpolation of recordings of a male speaker and a female speaker uttering the same word from an excellent database of American English Vowels (Hillenbrand *et al.*, 1995) but can be applied to any other

pair of mObjects such as recordings of different emotional expressions from the same speaker.

○ Generating the average of two mObjects

To generate an average of the mObject1 (Figure 2A) and mObject2 (Figure 2C), a rate of 0.5 is set for each mObject, such that each the 5 mRates fields equals [0.5 0.5]. Running VoiceMultiMorph generates the new mObjectM which can then be synthesized, listened to, saved if necessary and inspected. Figure 2b shows the smoothed spectrogram of the 2-voice average. Note that since the Frequency anchors of the two original mObjects in this example correspond to frequencies of their first 3 formants, the Frequency anchors of the newly generated mObjectM also provide its 3 formant frequencies.

○ Generating a continuum between two sounds

To generate a continuum between the two sounds, variable rates are set using a parameter k: mRates=[k 1-k], with k between 0 and 1. k represents the weight of the first mObject, such that a value of k=1 (mRates=[1 0]) corresponds to the first mObject, a value of k=0 (mRates=[0 1]) corresponds to the second mObject, and values between 0 and 1 to interpolation between the 2 mObjects. Parameter k can be set at regular intervals between 0 and 1 to generate a series of mObjects regularly spaced between the 2 original mObjects. For instance k=[1:-0.1:0] will generate a series of 11 mObjects in 10% steps between the two original mObjects. Or k=[0.95:-0.15:0.05] will generate 7 mObjects in 15 % steps between a 95%/5% and a 5%/95% morphs – such that each sound of the continuum contains some part of the two original sounds (Figure 2D).

Such continua are particularly interesting as they provide series of stimuli with equal acoustical distance between any two consecutive stimuli and can thus be used to dissociate perceptual aspects of a particular voice dimension (here, gender) from the unavoidable physical differences between the stimuli. For instance, the same 15% acoustical difference between 2 consecutive gender morphs has very different perceptual effects depending on whether they are towards the extremes of the continuum, or towards the middle of the continuum where it leads to marked changes in perceived gender (Charest *et al.*, 2013; Pernet & Belin, 2012).

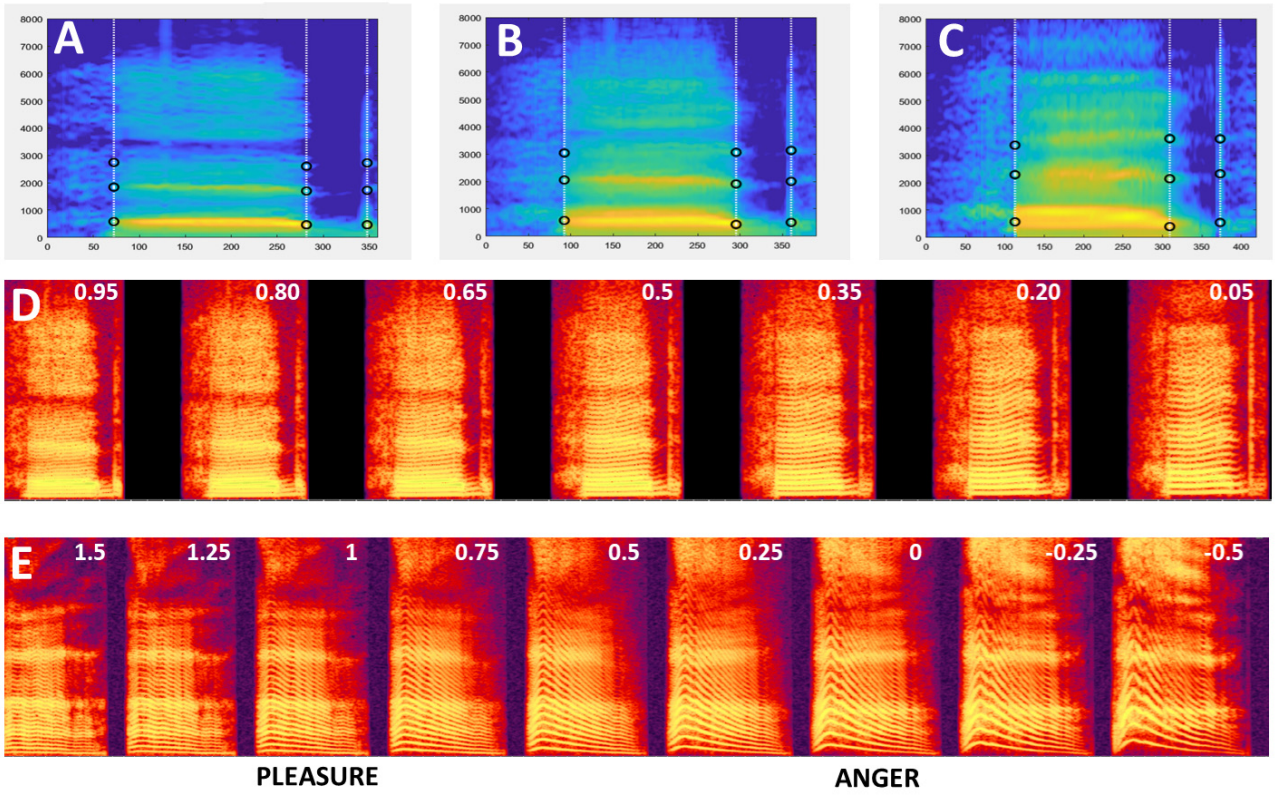

Figure 2. Morphing 2 stimuli. A–C. Smoothed spectrograms with Time and Frequency anchors of 3 mObjects. **A:** M48eh. **C:** W33eh. Note the similar definition of Time and Frequency anchors. **B**: Average of M48eh and W33eh. **D.** Spectrogram of the continuum generated between M48eh and W33eh using values of k =[0.95:-0.15:0.05] . Values of k are indicated at the top of each spectrogram. The leftmost stimulus is a 95%/5% mix of M48eh and W33eh, decreasing/increasing in steps of 15% until the rightmost stimulus – 5%/95% M48eh and 95% W33eh. The middle stimulus is the 50%/50% average shown in **B**. Any pair of consecutive stimuli is separated by the same acoustical distance at any point of the continuum: 15% of the M48eh-W33eh distance. **E.** Pleasure-Anger continuum including caricatures generated using values of k=[1.5:-.25:-0.5].

○ Caricaturing

In the continuum example of Figure 2D mRates kept values between 0 and 1, which makes intuitive sense but is not mathematically required. Indeed, while mRates values between 0 and 1 correspond to stimuli generated between mObject1 and mObject2 in the line between the two in multidimensional voice space, values below 0 and above 1 correspond to points beyond the two mObjects on that same line; i.e., caricatures (Calder *et al.*, 2000; Whiting *et al.*, 2020).

For instance, mRates = [1.5 -0.5] correspond to a point on the line at 50% of the mObject1-mObject2 distance on the side of mObject1 – in the example above, a voice exaggerating gender difference towards the male voice – while mRates=[-0.5 1.5] will produce and exaggeration of the same magnitude but towards the female voice. So by using mRates=[k 1-k] and k extending beyond 0 and 1 the continuum will be extended on both sides with increasingly strong caricatures. Figure 2E shows the spectrogram of an emotion continuum generated using k=[1.5:-0.1:-0.5] between a pleasure (6_pleasure.wav) and an angry (6_anger.wav) expression from an actor of the Montreal Affective Voices (Belin *et al.*, 2008). Note that using mRates values too negative or too much above 1 will likely result in unnatural sounding stimuli...

○ Parameter-specific morphing

In the examples above all 5 fields of the mRates structure were set to the same values. But it can be desirable in some cases to set some fields to different values than other, what has been called 'parameter-specific morphing' (Skuk *et al.*, 2020). For instance, all stimuli of the continuum could be set to the same duration – the average duration of mObjects 1 and 2 – by using mRates.duration=[0.5 0.5] while mRates in the other fields vary with k. Or, the anchorFrequency field could be set to [0.5 0.5] while the others vary with k, to generate a continuum of stimuli all having the formant frequencies of the mObject1-mObject2 average.

Morphing N mObjects A significant feature of VoiceMultiMorph is that it treats morphing N>2 mObjects in exactly the same way as for 2 mObjects. Simply, there are N dimensions instead of 2 to the multidimensional voice space defined by the mObjects, and mRates fields consist of a vector of N values (still adding up to 1 for each field) that can be viewed as the coordinates of the new mObjectM in N-dimensional voice space. Rather than averaging in turn all stimuli in successive steps, the user simply defines the desired coordinates in that space, and performs a single generation of mObjectM and resynthesis no matter how many stimuli are combined. In the examples below we consider 16 mObjects generated from recordings of the word 'Hello' spoken by 16 female speakers (Figure 3A) as part of a study on voice personality impressions (McAleer *et al.*, 2014). The first step consists in loading the 16 mObjects in a cell array of structures.

○ Generating the N-average

As for 2 mObjects, to generate the average of N mObjects the mRates vectors are all set to the same value of 1/N, here mRates=[1/16 … 1/16] or ones(1,16)/16. The mObjectM generated by these mRates values can be viewed as the center of N-dimensional voice space. It has average f0 contour and duration; it has average aperiodicity map, resulting in a very smooth spectrogram (Figure 3B) with high harmonics-to-noise ratio. Such an average stimulus is typically perceived as more attractive than individual voices (Bruckert *et al.*, 2010) and elicits lower activity in the TVAs than individual voices (Bestelmeyer *et al.*, 2011; Latinus *et al.*, 2013)– potentially reflecting a norm-based coding mechanism relative to an internal prototype.

○ Morphing with the N-average

It can be interesting to interpolate between a single mObject and the N-Average: studies in the face perception domain use this strategy to modulate the 'identity strength' of the faces (Leopold *et al.*, 2001). To do so, one strategy could be to first generate the mObject corresponding to the N-Average, then generate the continuum between the 2 mObjects as in the example above. Another, better strategy, that skips an unnecessary step, is to specify directly the mRates of the desired continuum trajectory in N-dimensional space.

With k=[1.5:-0.1:-0.5] as in the example above, to interpolate between the N-Average and mObject1, using mRates = [k*ones(1,16)/16] + [(1-k)*[1 zeros(1,15)]] or ([1-k+k/16 k/16 k/16… k/16] will generate stimuli on the line in multidimensional space going through mObject1 and the N-Average. K=1 corresponds to the average; k=0 to mObject 1. Values of k between 0 and 1 correspond to stimuli between the N-Average and mObject1 ('anti-caricatures'); values of k below 0 correspond to caricatures of mObject1 relative to the average. Values of k above 1 correspond to 'anti-voices', i.e. stimuli on the other side of the average relative to mObject1 in voice space, with opposite characteristics, e.g., higher than average f0 if mObject1 has lower f0 than average, etc. (Figure 3C)

Generating continua with the N-average can also be used for instance to take advantage of its high harmonics-to-noise ratio and generate 'smoother' or 'rougher' versions of mObjects. By using mRates values of [1 0 0 … 0] for all mRates fields, except for mRates.aperiodicity=[k*ones(1,16)/16] + [(1-k)*[1 zeros(1,15)]] stimuli will keep all aspects of mObject 1 but for the degree of aperiodic noise, and moving it closer to the average will generate a version of mObject1 with less aperiodicity, which listeners find more attractive (Bruckert *et al.*, 2010)

○ Generating random stimuli

The N mObjects can be viewed as the axes of an N-dimensional voice space they define. While all points of that space likely do not correspond to naturalistic voice stimuli – depending of course on the choice of the N mObjects – many points in the region around the average likely do. The N mObjects thus constitute a generative space that can be used to synthesize any point of the space. By applying random weights to each mObject in the mRates (still ensuring the total adds up to 1 per field), it is possible to sample the voice space by generating

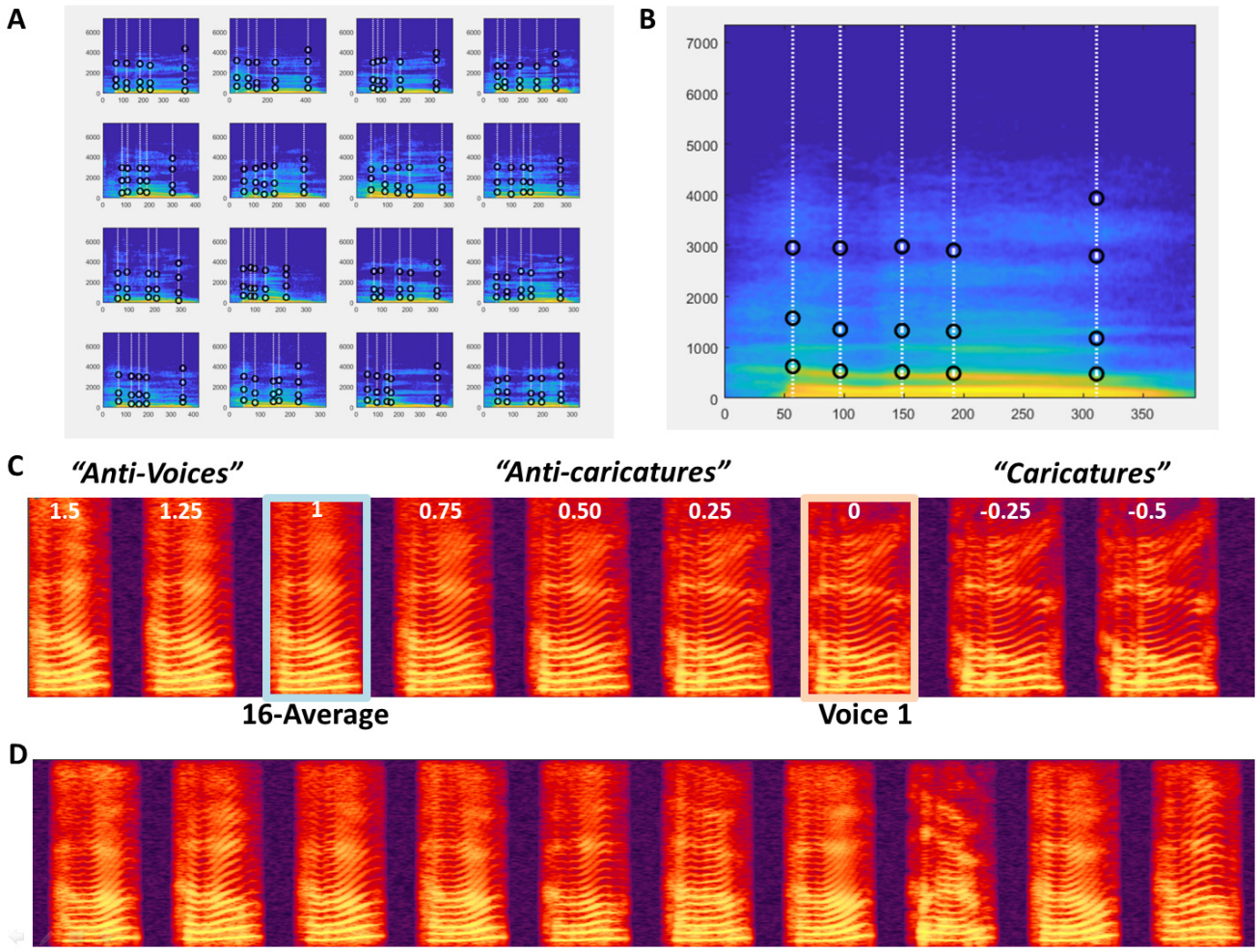

Figure 3. Morphing 16 mObjects. A. 16 mObjects of the word Hello spoken by 16 female speakers. Note their similar definition of the Time and Frequency anchors for combined morphing. **B.** MObjectM corresponding to the average of the 16 mObjects. **C.** Spectrogram of Continuum between Voice1 and the 16-Average with negative values of k for Caricatures and values of k greater than 1 for 'Antivoices' (Latinus & Belin, 2011). **D.** 10 random 'Hellos' generated using random mRates.

an arbitrary number of random stimuli. The example in the STRAIGHTMORPH_tutorial.m script applies random weights, with a coefficient that modulates the radius of the hyper-sphere around the average in which stimuli are generated, in order to generate an arbitrary number of random 'Hellos' and keeps the same weights across the 5 fields of the mRates.

Morphing sounds other than human voice

Finally, STRAIGHTMORPH can be used to morph stimuli other than human voices, provided there is a relatively clear f0. For instance, it is possible to generate a continuum between a human voice and a note played by a musical instrument (Belizaire *et al.*, 2007). Several groups have applied STRAIGHT to manipulation and morphing of vocalization from nonhuman primates (Chakladar *et al.*, 2008; Furuyama *et al.*, 2017). ExtractmObject will work well on these vocalizations provided the periodic content is strong enough but may require adjustment of the default values for stimuli with e.g. high f0. Figure 4A shows an example of identity continuum for macaque 'Coo' vocalizations similar to that of Figure 3C. For stimuli with very high f0, such as marmoset vocalizations, good results can be obtained by first transposing stimuli in the speech range by adjusting the sampling rate, then perform morphing, then readjusting the sampling rate of the generated stimuli (Figure 4B).

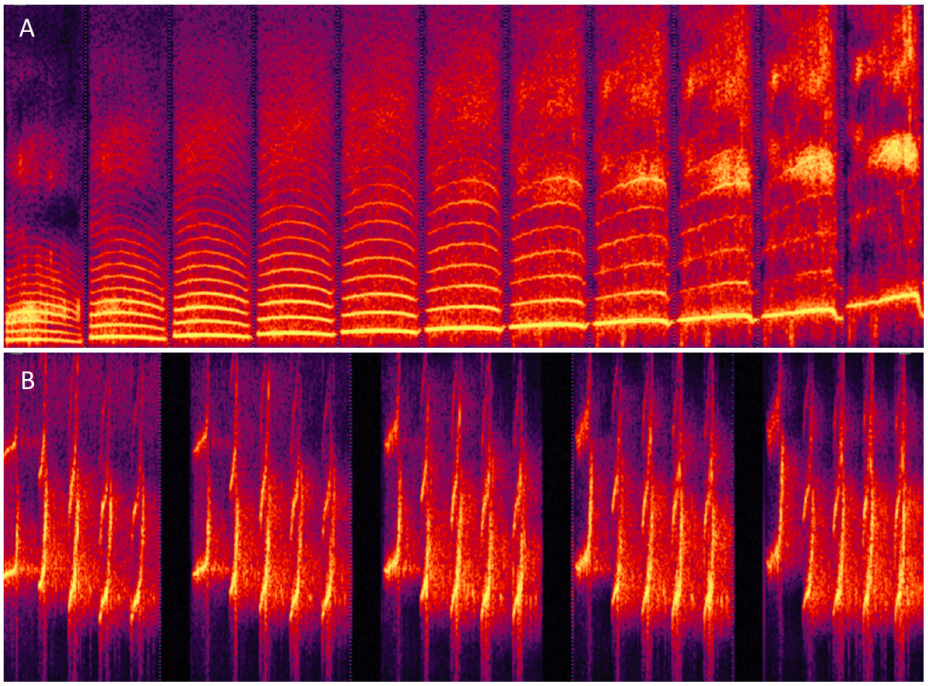

Figure 4. Morphing non-human primate vocalizations. A. Continuum between an individual macaque Coo and the average of 16 Coos using values of k similar to Figure 3C. **B.** Continuum between Twitter calls from 2 individual marmoset monkeys.

Data availability

STRAIGHTMORPH source code, documents and tutorial examples are available at: https://github.com/pascalbelin/STRAIGHTMORPH/ and https://doi.org/10.5281/zenodo.12097810 (Belin & Kawahara, 2024) under Creative Commons Attribution 4.0 International license.

Software availability statement

- Source code available from: https://github.com/pascalbelin/STRAIGHTMORPH/
- Archived software available from: DOI 10.5281/zenodo.12097810
- License: STRAIGHTMORPH is under Creative Commons Attribution 4.0 International licence

Acknowledgements

We acknowledge the support of UK's Economic and Social Research Council/Medical Research Council grant (RES-060-25-0010), Biotechnology and Biological Sciences Research Council grant (BBE0039581), French Fondation pour la Recherche Médicale (AJE201214), French Agence Nationale de la Recherche (ANR-16-CE37-0011-01 PRIMAVOICE; ANR-16-CONV-0002 Institute for Language, Communication and the Brain; ANR-11-LABX-0036 Brain and Language Research Institute), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (COVOPRIM788240). The legacy STRAIGHT and its open sourcing was supported by the CREST program of the Japan Science and Technology Agency, the e-Society program of the Ministry of Education, Culture, Sports, Science, and Technology Japan, and the Advanced Telecommunications Research Institute International in Kyoto Japan.

References

Belin P, Boehme B, McAleer P: **The sound of trustworthiness: acoustic-based modulation of perceived voice personality.** *PLoS One.* 2017; **12**(10): e0185651.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Belin P, Fillion-Bilodeau S, Gosselin F: **The "montreal affective voices": a validated set of nonverbal affect bursts for research on auditory affective processing.** *Behav Res Methods.* 2008; **40**(2): 531–539.
PubMed Abstract | **Publisher Full Text**

Belin P, Kawahara H: **Straightmorph: a voice morphing tool for research in voice communication sciences.** *Zenodo.* 2024.
http://www.doi.org/10.1101/2024.06.04.597359

Belizaire G, Fillion-Bilodeau S, Chartrand JP, *et al.*: **Cerebral response to 'voiceness': a functional magnetic resonance imaging study.** *Neuroreport.* 2007; **18**(1): 29–33.
PubMed Abstract | **Publisher Full Text**

Bestelmeyer PEG, Kotz SA, Belin P: **Effects of emotional valence and arousal on the voice perception network.** *Soc Cogn Affect Neurosci.* 2017; **12**(8): 1351–1358.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Bestelmeyer PEG, Latinus M, Bruckert L, *et al.*: **Implicitly perceived vocal attractiveness modulates prefrontal cortex activity.** *Cereb Cortex.* 2011; **22**(6): 1263–1270.
PubMed Abstract | **Publisher Full Text**

Bestelmeyer PEG, Maurage P, Rouger J, *et al.*, **Adaptation to vocal expressions reveals multistep perception of auditory emotion.** *J Neurosci.* 2014; **34**(24): 8098–8105.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Bruckert L, Bestelmeyer P, Latinus M, *et al.*: **Vocal attractiveness increases by averaging.** *Curr Biol.* 2010; **20**(2): 116–120.
PubMed Abstract | **Publisher Full Text**

Calder AJ, Rowland D, Young AW, *et al.*: **Caricaturing facial expressions.** *Cognition.* 2000; **76**(2): 105–146.
PubMed Abstract | **Publisher Full Text**

Chakladar S, Logothetis NK, Petkov CI: **Morphing rhesus monkey vocalizations.** *J Neurosci Methods.* 2008; **170**(1): 45–55.
PubMed Abstract | **Publisher Full Text**

Charest I, Pernet C, Latinus M, *et al.*: **Cerebral processing of voice gender studied using a continuous carryover FMRI design.** *Cereb Cortex.* 2013; **23**(4): 958–966.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Furuyama T, Kobayasi KI, Riquimaroux H: **Acoustic characteristics used by Japanese macaques for individual discrimination.** *J Exp Biol.* 2017; **220**(Pt 19): 3571–3578.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Giordano BL, Whiting C, Kriegeskorte N, *et al.*: **The representational dynamics of perceived voice emotions evolve from categories to dimensions.** *Nat Hum Behav.* 2021; **5**(9): 1203–1213.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hillenbrand J, Getty LA, Clark MJ, *et al.*: **Acoustic characteristics of American english vowels.** *J Acoust Soc Am.* 1995; **97**(5 Pt 1): 3099–3111.
PubMed Abstract | **Publisher Full Text**

Kawahara H, Matsui H: **Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation.** *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.* 2003; 256–259.
Publisher Full Text

Kawahara H, Morise M: **Interactive tools for making vocoder-based signal processing accessible: flexible manipulation of speech attributes for explorational research and education.** *Acoust Sci Tech.* 2024; **45**(1): 48–51.
Publisher Full Text

Koyano KW, Jones AP, McMahon DBT, *et al.*: **Dynamic suppression of average facial structure shapes neural tuning in three macaque face patches.** *Curr Biol.* 2021; **31**(1): 1–12.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Latinus M, Belin P: **Anti-voice adaptation suggests prototype-based coding of voice identity.** *Front Psychol.* 2011; **2**: 175.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Latinus M, McAleer P, Bestelmeyer PEG, *et al.*: **Norm-based coding of voice identity in human auditory cortex.** *Curr Biol.* 2013; **23**(12): 1075–1080.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Leopold DA, Bondar IV, Giese MA: **Norm-based face encoding by single neurons in the monkey inferotemporal cortex.** *Nature.* 2006; **442**(7102): 572–575.
PubMed Abstract | **Publisher Full Text**

Leopold DA, O'Toole AJ, Vetter T, *et al.*: **Prototype-referenced shape encoding revealed by high-level aftereffects.** *Nat Neurosci.* 2001; **4**(1): 89–94.
PubMed Abstract | **Publisher Full Text**

McAleer P, Todorov A, Belin P: **How do you say 'hello'? Personality impressions from brief novel voices.** *PLoS One.* 2014; **9**(3): e90779.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Morris JS, Frith CD, Perrett DI, *et al.*: **A differential neural response in the human amygdala to fearful and happy facial expressions.** *Nature.* 1996; **383**(6603): 812–815.
PubMed Abstract | **Publisher Full Text**

Nussbaum C, von Eiff CI, Skuk VG, *et al.*: **Vocal emotion adaptation aftereffects within and across speaker genders: roles of timbre and fundamental frequency.** *Cognition.* 2022; **219**: 104967.
PubMed Abstract | **Publisher Full Text**

Pernet CR, Belin P: **The role of pitch and timbre in voice gender categorization.** *Front Psychol.* 2012; **3**: 23.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rotshtein P, Henson RNA, Treves A, *et al.*: **Morphing marilyn into maggie dissociates physical and identity face representations in the brain.** *Nat Neurosci.* 2005; **8**(1): 107–113.
PubMed Abstract | **Publisher Full Text**

Skuk VG, Kirchen L, Oberhoffner T, *et al.*: **Parameter-specific morphing reveals contributions of timbre and fundamental frequency cues to the perception of voice gender and age in cochlear implant users.** *J Speech Lang Hear Res.* 2020; **63**(9): 3155–3175.
PubMed Abstract | **Publisher Full Text**

von Eiff CI, Fruhholz S, Korth D, *et al.*: **Crossmodal benefits to vocal emotion perception in cochlear implant users.** *iScience.* 2022; **25**(12): 105711.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Whiting CM, Kotz SA, Gross J, *et al.*: **The perception of caricatured emotion in voice.** *Cognition.* 2020; **200**: 104249.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔ ✔ ?

Version 2

Reviewer Report 03 March 2025

https://doi.org/10.21956/openreseurope.20747.r49993

© **2025 Miller L.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Lee M. Miller
University of California, Davis, USA

This report introduces a software tool for parametric acoustical morphing among voices with different identity or expression: STRAIGHTMORPH. It is built on Legacy-STRAIGHT, a well established and highly regarded speech analysis-synthesis (vocoder) package developed by the second author. The morphing capability of Legacy-STRAIGHT has existed for many years, but there lacked a user friendly interface to take full advantage of it. STRAIGHTMORPH supports parametric morphing along five basic acoustic dimensions (F0, spectral amplitude, aperiodicity, time, and frequency) which capture many of the attributes that convey perceptual distinctions among voices (or other complex natural sounds with pitch such as nonhuman primate vocalizations or music). Morphing along one or more of these dimensions allows the user to produce high quality voice continua, averages, and caricatures. Importantly, this allows one to dissociate perceptual changes from well-controlled acoustical changes; consequently, a primary application area is neurobehavioral studies of voice perception. This report gives step-by-step instructions on how to use the helpful GUI in a variety of use cases. It does not - and need not - unpack the technical foundations of the analysis-synthesis in Tandem-STRAIGHT, as this is covered in numerous prior publications. All code including a tutorial function are freely available on github. STRAIGHTMORPH offers a tremendously valuable, powerful, and easy-to-use tool for a wide variety of voice-acoustic researchers.

SPECIFIC COMMENTS (page numbers refer to the pdf)

MAJOR:

- Not everyone is adept at reading spectrograms. So that readers can appreciate what the morphing is actually doing, please embed the audio for the examples shown. This would increase the clarity and impact of the report.
- A brief explanation and justification of using Legacy-STRAIGHT, as opposed to TANDEM-STRAIGHT (or WORLD, for that matter), would be welcome. Why not use TANDEM-STRAIGHT, which was developed as an improved successor to Legacy-STRAIGHT?
- p5. I find the term “morphing rate” to be very confusing, because it is in no straightforward way a “rate”. Since the values indicate the weighting of each mObject in a morph for a given

parameter (F0, spectralamplitude, etc.), a more accurate term would be “morphing weight”.
- A brief passage on practical limitations would be helpful. How different can utterances be before the morphing typically fails? How much correspondence must there be between the time and frequency anchors? What vocal attributes are NOT captured well or readily parameterizable in an mObject?

MINOR:
- p3. “a similar implement” –> “a similar implementation”
- p4. “Object.aperiodicityIndex containing a matrix of similar size as the spectrogram” Shouldn’t “similar” instead be “identical”?
- p3-4. Please define the terms “mObject” and “anchor” briefly upon first use in main text, rather than waiting for the full explication later.
- p4: “spoken a female” -> “spoken by a female”
- “it can be first be” -> “it can first be”
- Figure 1 legend. C: Please define blue and green lines (presumably the manual f0 range restriction).
- p5 “cell array of structure” -> cell array of structures”
- Figure 2 legend. It would help to explain briefly why the representations differ between A-C and D-E. For instance “A-C. Smoothed spectrograms highlight coarse features such as formants.” “D-E. Unsmoothed spectrograms highlight fine structure such as harmonics and their spacing (F0)”.
- p7. “unnatural sounding stimuli…” replace ellipsis with period.
- p7. define “TVAs”. Alternately, leave this entire sentence out, as the deep dive seems more fitting for a Discussion than a technical explication.
- p7. The term “anti-voices” makes sense, but it’s not clear why a morph between the N-average and a given mObject is an “anti-caricature”.
- p7. “While all points of that space likely do not correspond” should read “While not all points of that space likely correspond”, or perhaps clearer, “While it is unlikely that all points of that space correspond”.
- p7 Parameter-specific morphing. You say “mRates.duration”, but on p5 the only 5 fields of mRates are listed as: F0, spectralamplitude, aperiodicity, time, and frequency. Same for “frequency” vs “anchorFrequency”. Please use the same field/parameter names throughout.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neurobiology of speech perception

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 27 January 2025

https://doi.org/10.21956/openreseurope.20747.r49705

© 2025 Schweinberger S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Stefan Schweinberger
University of Jena, Jena, Germany

This is a crisp and clearly written paper which describes the voice morphing tool STRAIGHTMORPH – which is based on the long-standing work of the second author of the paper, and which serves the community of voice communication scientist by making high-quality software more accessible.

This is an excellent initiative, and with its high accessibility and fine documentation, there can be no doubt that STRAIGHTMORPH will be a useful resource for the growing community of voice researchers.

I only have a few minor points regarding this already revised version:

- In the first paragraph, the authors emphasize how it is important to parametrically manipulate the physical characteristics of complex naturalistic stimuli. In this context, while I believe it is adequate if stimulus examples for the software package are kept relatively simple (e.g., vowels, single words, or primate vocalizations, a few comments on how STRAIGHTMORPH can be used to morph complex naturalistic vocalisations (e.g., sentences) would be welcome, and these might enhance the impact of the software tool.
- The Section Legacy STRAIGHT is highly welcome. Two relevant references might be worth adding here: Specifically, I´m referring to the Methods Chapter “Voice Morphing” by Skuk and Kawahara (in S. Frühholz and P. Belin (Eds.): *The Oxford Handbook of Voice Perception Chapter 31, pp. 685-706*), which provides a relatively extensive discussion of voice morphing technology, and also to an early empirical paper from 2008 (https://doi.org/10.1016/j.cub.2008.04.015) which could be added to the examples for how the method generated exciting results. [Schweinberger SR.et.al.2008 - Ref 1]
- At this point on p.3, please also change “an growing number of researchers” to “a growing number of researchers”
- Similarly on p4, “an time-frequency map” should read “a time-frequency map”.

- On P.7, 2nd paragraph “Note that using mRates values too negative or too much above 1 will likely result in unnatural sounding stimuli…”: My recommendation is to omit the multiple dots at the end. If possible, be more precise as to the ways in which this may result to unnatural sounding stimuli.

P.7: a blank seems needed after the closed parenthesis in “…Latinus et al., 2013)– potentially”.

References

1. Schweinberger SR, Casper C, Hauthal N, Kaufmann JM, et al.: Auditory adaptation in voice perception.*Curr Biol*. 2008; **18** (9): 684-8 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Voice Perception, Person Perception, Social Interaction, Social Neuroscience

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 January 2025

https://doi.org/10.21956/openreseurope.20747.r49651

© 2025 Lavan N. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nadine Lavan
Queen Mary University of London, London, England, UK

The authors have addressed my initial concern in full. I am thus happy to recommend for this manuscript to be approved.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Voice perception

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 January 2025

https://doi.org/10.21956/openreseurope.20747.r49723

© 2025 Anikin A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Andrey Anikin
Univ. of Lund, Lund,, Sweden

Belin and Kawahara present a non-technical, practically oriented description of the well-known STRAIGHT algorithm for manipulating and morphing voice recordings, which is now made freely available as a collection of Matlab functions. This is a highly specialized tool, but it has nevertheless proved very useful for research on voice perception, as the authors clearly demonstrate in the text. The availability of this user guide will no doubt make the tool more accessible and help to popularize it further. I have nothing to criticize in the paper itself. Merely as a suggestion for future work, I would consider using an open-source framework (Python, R, Julia, ...) instead of Matlab. Another obvious extension is to add functionality for modifying the source and/or filter after their separation. The filter can be arbitrarily warped (cf. phase vocoders like CLEESE, DAVID etc.), and the source can easily be resynthesized prior to being recombined with the filter for ultimate control (cf. soundgen).

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human nonverbal vocalizations, voice analysis and synthesis, auditory

attention

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 10 September 2024

https://doi.org/10.21956/openreseurope.19511.r42894

© 2024 Liu Y. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Yi-Wen Liu
[1] National Tsing Hua University, Hsinch, Taiwan
[2] National Tsing Hua University, Hsinch, Taiwan

As also indicated by the first reviewer, this article is written with clarity and users should be able to follow the tutorials and start using the tool. The applications in Fig. 3 (16-voice morphing) and Fig. 4 (non-primate vocalization) are particularly intriguing to me. With the clever design of user interface which allows user to work intuitively with mObjects, I believe this tool will impact the field and encourage more researchers to try morphing voices for their projects.

Below are a few editorial suggestions (for the printed version):

1. In the paragraph Legacy STRAIGHT: the apostrophe seem to be misplaced before the word spectrogram ('spectrogram)

2. to try and correct -> "to try to correct", or "to correct"?

3. The term "time axes" is a bit puzzling to me. Does it mean "time location" or simply "time anchor"? I currently assume that a "time axis" is one of the white vertical dotted lines on Fig. 1D and 1E.

4. "likely result in unnatural sounding stimuli..." -> remove the trailing periods "..."

5. In the paragraph "Generating the N-average", what does the abbreviation TVA stand for?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: audio and speech signal processing, hearing science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Dec 2024

Pascal Belin

Many thanks for your positive evaluation.

Competing Interests: No competing interests were disclosed.

Reviewer Report 31 July 2024

https://doi.org/10.21956/openreseurope.19511.r42639

© 2024 Lavan N. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Nadine Lavan
[1] Queen Mary University of London, London, England, UK
[2] Queen Mary University of London, London, England, UK

This is a clearly written article, describing the voice morphing tool STRAIGHTMORPH.

From my reading of the paper, all the steps required to morph voices (and other sounds) are described clearly and in an easy-to-follow way. The toolbox is easily accessible via links provided in the paper. Overall, STRAIGHTMORPH is a useful resource for the voice research community.

I have therefore only one comment: Voice morphing has been used quite widely (in the form of parameter-specific morphing or otherwise) to answer many different research questions. The authors across the manuscript list a number of different papers that have used morphing (e.g. in

the section "legacy STRAIGHT") and even provide some details on a small number of studies that have used morphing (see "Morphing Examples"). I however wonder whether a reader who has not encountered voice morphing before as a technique will be able to get a good idea of how versatile voice morphing can indeed be. I believe that adding some additional examples/use cases of how voice morphing has been used before in a bit more detail in the earlier sections might make this paper more impactful overall.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Voice perception

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 11 Dec 2024

Pascal Belin

Thank you for your positive evaluation and your valuable suggestion: readers are now referred to voice morphing examples beyond those described in the manuscript.

Competing Interests: No competing interests were disclosed.