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

P Belin[1], H Kawahara[2]

1. Institut de Neurosciences de la Timone, UMR7289 CNRS & Aix-Marseille Université, Marseille, France.

2. Wakayama University, Wakayama, Japan.

Corresponding authors: Pascal Belin, Institut de Neurosciences de la Timone, 27 bd Jean Moulin, 13385 Marseille cedex 5, France. pascal.belin@univ-amu.fr ; and Hideki Kawahara, ADDRESS kawahara@wakayama-u.ac.jp

Abstract

The purpose of the present 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 to generate 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. Although STRAIGHTMORPH has been designed for human voices it can work well with other types of sounds such as non-human primate vocalizations.

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; D. A. Leopold, Bondar, & Giese, 2006; D.A. Leopold, O'Toole, Vetter, & Blanz, 2001; Morris et al., 1996; Rotshtein, Henson, Treves, Driver, & Dolan, 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, Pernet, Latinus, Crabbe, & Belin, 2013; Latinus, McAleer, Bestelmeyer, & Belin, 2013; Nussbaum, von Eiff, Skuk, & Schweinberger, 2022; Skuk et al., 2020 ; von Eiff, Fruhholz, Korth, Guntinas-Lichius, & Schweinberger, 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, Logothetis, & Petkov, 2008; Furuyama, Kobayasi, & Riquimaroux, 2017)

The purpose of the present 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 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

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

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) x 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 (Fig. 1B,C).

Upon calling the function a time-frequency graph is displayed showing different f0 candidates. 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.

(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 (Fig 1E).
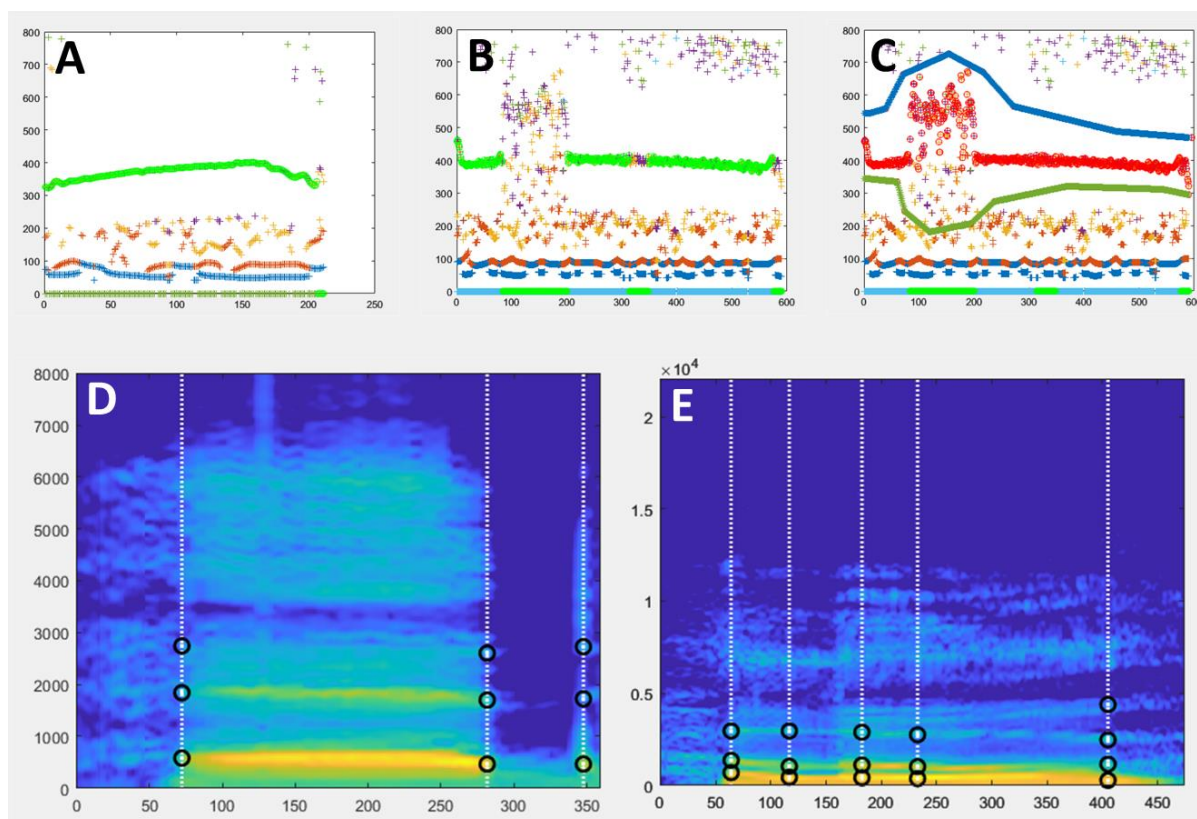

*Figure 1. mObject extraction. **A.** Example of successful automatic f0 extraction using ExtractMObject. The diagram (X axis: time (ms); Y axis: frequency(Hz)) shows identified f0 candidates as coloured circles. F0 candidates selected by the automatic procedure are highlighted in green. **B.** Example 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.*

Once an mObject has been generated, it can be first be inspected using the STRAIGHT function displayMObject (Code in Appendix lines xx). 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 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, Getty, Clark, & Wheeler, 1995) but can be applied to any other pair of mObjects such as recordings of two different emotional expressions from the same speaker.

o   Generating the average of two mObjects

To generate an average of the mObject1 (Fig. 2A) and mObject2 (Fig. 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.

o   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 <u>they provide series of stimuli with equal acoustical distance between any two consecutive stimuli</u> 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).
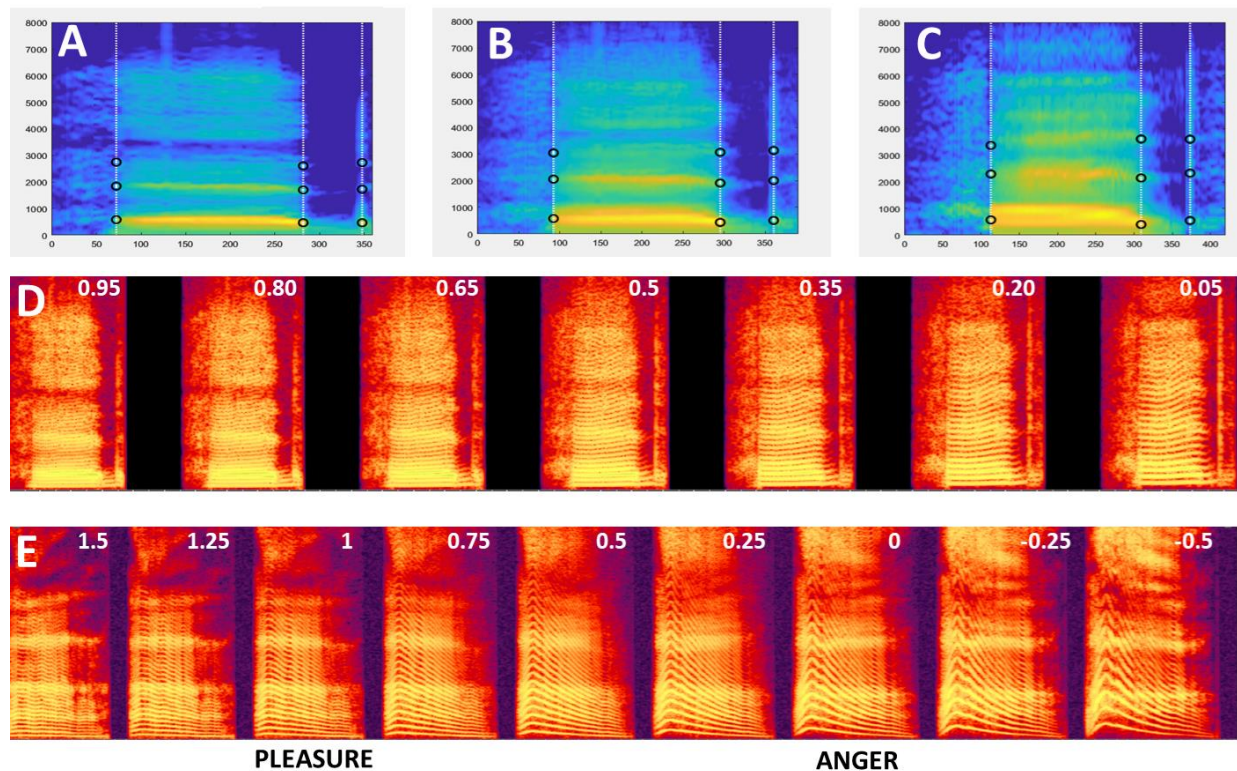

Figure 2. Morphing 2 stimuli. A-C. Smoothed spectrograms with Time and Frequency anchors of 3 Mobjects. A: M48eh. C: W33eh. Note the similar definition of Time and Frequency anchors. B: Average of M48eh and W33eh. D. Spectrogram of the continuum generated between M48eh and W33eh using values of k =[0.95:-0.15:0.05] . 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].

  o 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.

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 . Fig 3E 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, Fillion-Bilodeau, & Gosselin, 2008). Note that using mRates values too negative or too much above 1 will likely result in unnatural sounding stimuli...

  o 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 (Fig. 3A) as part of a study on voice personality impressions (McAleer, Todorov, & Belin, 2014). The first step consists in loading the 16 mObjects in a cell array of structures.

o   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 (Fig 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.

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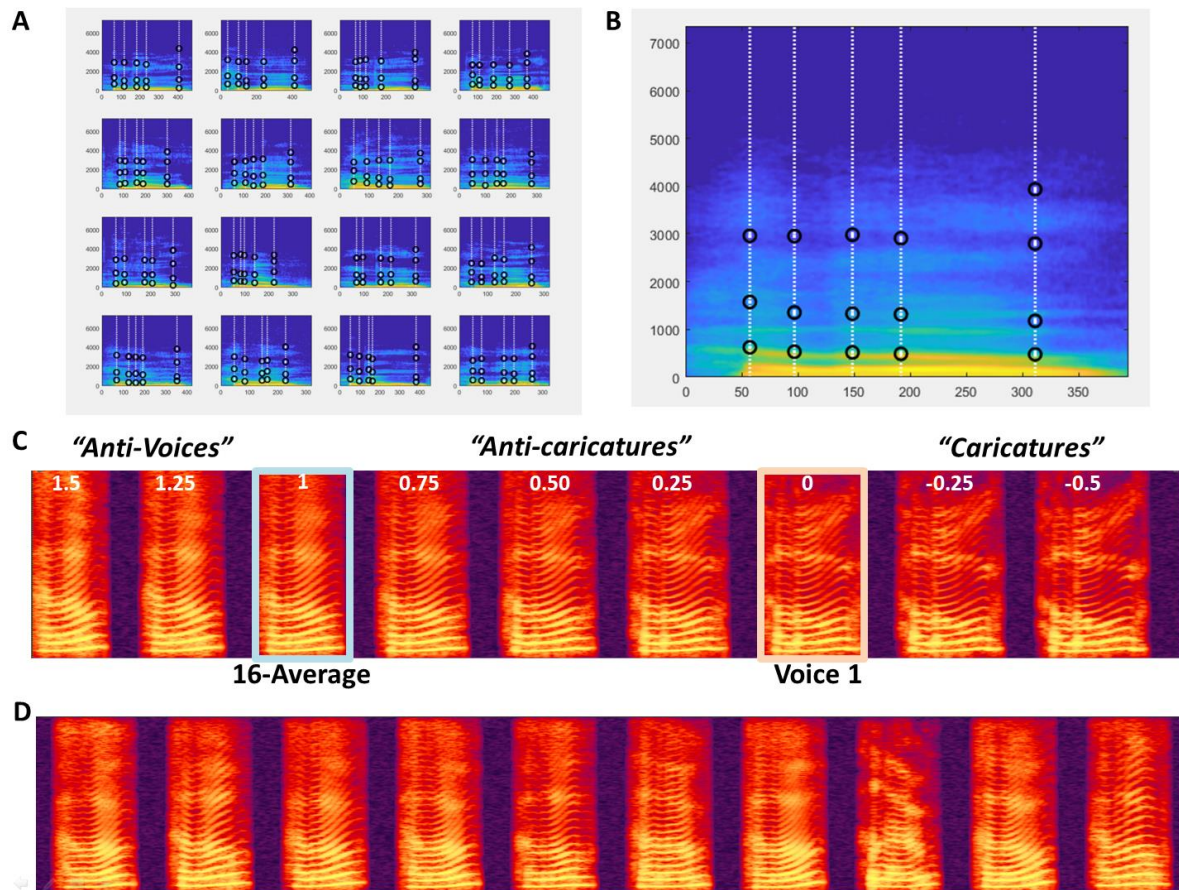
Figure 3. Morphing 16 mObjects. A. *16 mObjects of the word Hello spoken by 16 female speakers. Note their similar definition of the Time and Frequency anchors for combined morphing.* **B.** *mObjectM corresponding to the average of the 16 mObjects. C. Spectrogram of Continuum between Voice1 and the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

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 O 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)

o   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 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 hypersphere 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, Fillion-Bilodeau, Chartrand, Bertrand-Gauvin, & Belin, 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 Fig 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 (Fig. 4B).
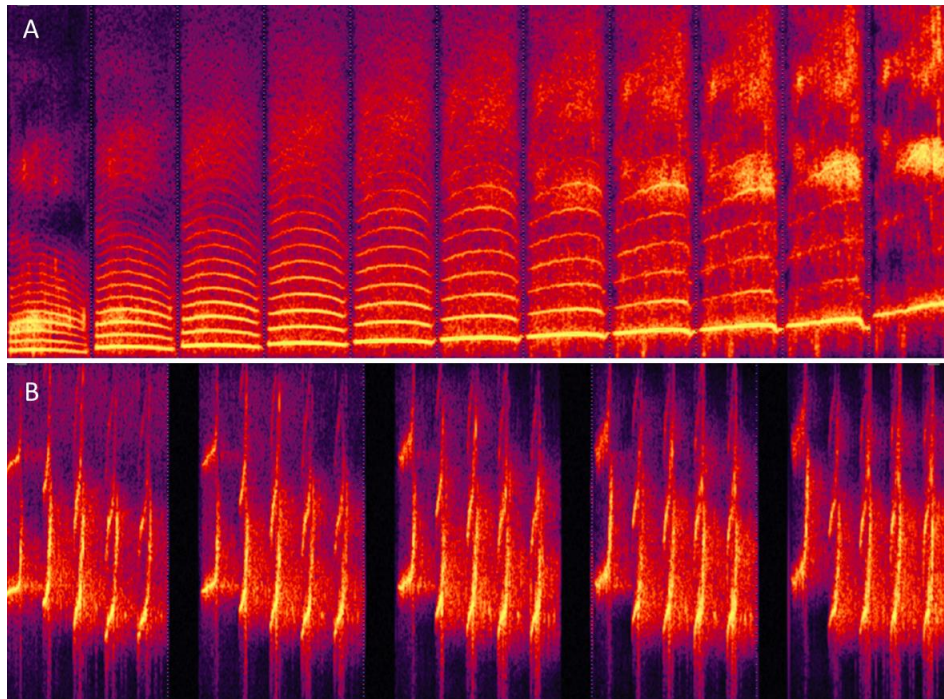

Figure 4. Morphing non-human primate vocalizations. A. Continuum between an individual macaque Coo and the average of 16 Coos using values of k similar to Fig. 3C. B. Continuum between Twitter calls from 2 individual marmoset monkeys.

ACKNOWLEDGEMENTS

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

(788240). HIDEKI?

REFERENCES

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The "Montreal Affective Voices": a validated set of nonverbal affect bursts for research on auditory affective processing. *Behav Brain Res, 40*, 531-539.

Belizaire, G., Fillion-Bilodeau, S., Chartrand, J. P., Bertrand-Gauvin, C., & Belin, P. (2007). Cerebral response to 'voiceness': a functional magnetic resonance imaging study. *Neuroreport, 18*(1), 29-33. doi:10.1097/WNR.0b013e32801227180000 1756-200701080-00006 [pii]

Bestelmeyer, P. E., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., & Belin, P. (2011). Implicitly Perceived Vocal Attractiveness Modulates Prefrontal Cortex Activity. *Cereb Cortex*. doi:bhr204 [pii]10.1093/cercor/bhr204

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., . . . Belin, P. (2010). Vocal attractiveness increases by averaging. *Curr Biol, 20*(2), 116-120. doi:S0960-9822(09)02049-1 [pii]10.1016/j.cub.2009.11.034

Chakladar, S., Logothetis, N. K., & Petkov, C. I. (2008). Morphing rhesus monkey vocalizations. *J Neurosci Methods, 170*(1), 45-55. doi:10.1016/j.jneumeth.2007.12.023

Charest, I., Pernet, C., Latinus, M., Crabbe, F., & Belin, P. (2013). Cerebral processing of voice gender studied using a continuous carryover FMRI design. *Cereb Cortex, 23*(4), 958-966. doi:10.1093/cercor/bhs090 [pii]

Furuyama, T., Kobayasi, K. I., & Riquimaroux, H. (2017). Acoustic characteristics used by Japanese macaques for individual discrimination. *J Exp Biol, 220*(Pt 19), 3571-3578. doi:10.1242/jeb.154765

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099-3111.

Kawahara, H., & Matsui, H. (2003). 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*, 256-259.

Kawahara, H., & Morise, M. (2024). Interactive tools for making vocoder-based signal processing accessible: Flexible manipulation of speech attributes for explorational research and education. *Acoust Sci & Tech, 45*(1).

Koyano, K. W., Jones, A. P., McMahon, D. B. T., Waidmann, E. N., Russ, B. E., & Leopold, D. A. (2021). Dynamic Suppression of Average Facial Structure Shapes Neural Tuning in Three Macaque Face Patches. *Curr Biol, 31*(1), 1-12 e15. doi:10.1016/j.cub.2020.09.070

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Front Psychol, 2*, 175. doi:10.3389/fpsyg.2011.00175

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Curr Biol, 23*(12), 1075-1080. doi:10.1016/j.cub.2013.04.055 S0960-9822(13)00496-X [pii]

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442*(7102), 572-575. Retrieved from <Go to ISI>://000239455900045

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat Neurosci, 4*, 89-94.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One, 9*(3), e90779. doi:10.1371/journal.pone.0090779 PONE-D-13-23379 [pii]

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383*, 812-815.

Nussbaum, C., von Eiff, C. I., Skuk, V. G., & Schweinberger, S. R. (2022). Vocal emotion adaptation aftereffects within and across speaker genders: Roles of timbre and fundamental frequency. *Cognition, 219*, 104967. doi:10.1016/j.cognition.2021.104967

Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Front Psychol, 3*, 23. doi:10.3389/fpsyg.2012.00023

Rotshtein, P., Henson, R. N., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci, 8*, 107-113.

Skuk, V. G., Kirchen, L., Oberhoffner, T., Guntinas-Lichius, O., Dobel, C., & Schweinberger, S. R. (2020). 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, 63*(9), 3155-3175. doi:10.1044/2020_JSLHR-20-00026

von Eiff, C. I., Fruhholz, S., Korth, D., Guntinas-Lichius, O., & Schweinberger, S. R. (2022). Crossmodal benefits to vocal emotion perception in cochlear implant users. *iScience, 25*(12), 105711. doi:10.1016/j.isci.2022.105711