

Generation of a vocal-tract MRI movie based on sparse sampling

Sadao Hiroya¹, Tatsuya Kitamura²

¹NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, Japan

²Faculty of Intelligence and Informatics, Konan University
8-9-1 Okamoto, Higashinada, Kobe, Japan

hiroya.sadao@lab.ntt.co.jp, t-kitamu@konan-u.ac.jp

Abstract. *We present a novel technique that can provide a high-quality vocal-tract MRI movie during speech production. The method uses MRI vocal-tract images at the central point of each phoneme and interpolation functions between adjacent phonemes obtained from electromagnetic articulographic (EMA) data. It is based on our finding that articulatory parameters are suitable for a sparse representation. Preliminary results showed that the quality of the obtained vocal-tract MRI movie is high compared to that for the previous technique. The method will be useful for constructing a large database of vocal-tract MRI movies and understanding speech production mechanisms.*

1. Introduction

The measurement of articulatory motion during speech production is important for investigating human speech production mechanisms. In the past, X-ray cinematography (Perkell, 1969), an X-ray microbeam system (Kiritani et al., 1975), and an ultrasound technique (Stone et al., 1983) have been used to measure motions of human articulatory organs. In the 1990's, an electromagnetic articulographic (EMA) system was widely used for speech production research (Perkell et al., 1992). This system can measure the tongue, lips, incisors, and velum at fairly high rates, but can only track receiver coils in the mid-sagittal plane attached to the speech articulators. The advantage of this technique is that the calculation of velocity and acceleration is more direct with flesh points. Currently, a measuring system that works in three dimensions is being developed and used (Kaburagi et al., 2005).

Recently, it has become possible to measure complete two- or three-dimensional vocal-tract motion during speech production by magnetic resonance imaging (MRI). One of the imaging methods is synchronized sampling with an external trigger (Masaki et al., 1999). This can generate a high-quality vocal-tract MRI movie, but the subject has to repeat the same sentence more than 100 times with small trial-to-trial variability. Another is a real-time MRI technique (Narayanan et al., 2004), but lower sampling rates are used as for measuring articulatory motion during speech production. Therefore, a technique that can provide a high-quality vocal-tract MRI movie and greatly reduce the number of sentence repetitions in the MRI scanner is required.

We have proposed a method for decomposing EMA-based mid-sagittal articulatory parameters into a set of temporally overlapped event functions and corresponding event vectors using non-negative temporal decomposition (NTD) (Hiroya, 2010) and found that articulatory motion during speech production can be represented by articulatory positions at the central point of each phoneme and adequate interpolation functions.

In this paper, we present a method for generating a vocal-tract MRI movie based on sparse sampling: vocal-tract MRI image sequences during speech production can be generated using MRI images at key frames and interpolation functions obtained by EMA measurements. To verify this validity, we investigate the articulatory properties obtained with EMA and MRI, and compare the proposed method with the previous technique.

2. Non-negative temporal decomposition

The original temporal decomposition (Atal, 1983) approximates the i th parameter $y_i(t)$ of time t to

$$\hat{y}_i(t) = \sum_{k=1}^m a_{i,k} \phi_k(t), \quad 1 \leq t \leq T, \quad 1 \leq i \leq p, \quad (1)$$

where $a_{i,k}$ is the k th event vector, $\phi_k(t)$ is the k th event function, p is the dimension of the parameter, T is the length of the parameter sequence, and m is the number of event functions. Later, Shiraki (2004) and Kim and Oh (1999) assumed that $\phi_k(t)$ is zero for $t < t_{k-1}$ and $t > t_{k+1}$ and that the sum of all event functions is one at any time t :

$$\hat{y}_i(t) = a_{i,k} \phi_k(t) + a_{i,k-1} \phi_{k-1}(t), \quad t_{k-1} \leq t \leq t_k, \quad \text{where } \phi_k(t) + \phi_{k-1}(t) = 1. \quad (2)$$

They claimed that the event functions $\phi_k(t)$ should be restricted to the range $[0, 1]$, because this restriction facilitates parameter modification. The event functions are determined by minimizing the least-squares error between $y_i(t)$ and $\hat{y}_i(t)$, but the obtained event functions are not restricted to the range $[0, 1]$. Thus, Kim clipped event functions at the range $[0, 1]$ [i.e. $\phi_k(t) = \min(1, \max(0, \phi_k(t)))$], but this resulted in increasing the estimation errors.

It is difficult to determine the event functions that are restricted to the range $[0, 1]$ by the least-squares method without a clipping. To overcome this difficulty, we have proposed a method for decomposition of speech parameters into a set of temporally overlapped event functions $\phi_k(t)$ that are restricted to the range $[0, 1]$ and corresponding event vectors $a_{i,k}$ using NTD (Hiroya, 2010). The point is that NTD can determine the interpolation functions and the key frames based on a combination of non-negative matrix factorization (NMF) (Lee and Seung, 1999) and dynamic programming (DP). We consider minimizing the following cost function by the NMF algorithm:

$$\sum_{k=2}^m \sum_{t=t_{k-1}}^{t_k} \sum_{i=1}^p (y_i(t) - a_{i,k} \phi_k(t) - a_{i,k-1} \phi_{k-1}(t))^2 + \alpha \sum_{k=2}^m \sum_{t=t_{k-1}}^{t_k} (\phi_k(t) + \phi_{k-1}(t) - 1)^2, \quad (3)$$

where $1 = t_1 < t_2 < \dots < t_m = T$ and α is the weight. In line with a previous idea

(Virtanen, 2007), we can obtain the multiplicative update rule for the event function:

$$\phi_k(t) \leftarrow \frac{\sum_{i=1}^p a_{i,k} y_i(t) + \alpha}{\sum_{i=1}^p (a_{i,k-1} a_{i,k} \phi_{k-1}(t) + a_{i,k}^2 \phi_k(t)) + \alpha (\phi_{k-1}(t) + \phi_k(t))} \phi_k(t) \quad (4)$$

$$\phi_{k-1}(t) \leftarrow \frac{\sum_{i=1}^p a_{i,k-1} y_i(t) + \alpha}{\sum_{i=1}^p (a_{i,k-1} a_{i,k} \phi_k(t) + a_{i,k-1}^2 \phi_{k-1}(t)) + \alpha (\phi_{k-1}(t) + \phi_k(t))} \phi_{k-1}(t) \quad (5)$$

The distortion $d(y(t), \hat{y}(t))$ of the cost function for each interval $t_{k-1} \leq t \leq t_k$ only depends on time t_{k-1} and t_k . Therefore, the event timing

$$t_k = \arg \min_{t_2, \dots, t_{m-1}} \sum_{t=1}^T d(y(t), \hat{y}(t)), \quad 2 \leq k \leq m-1 \quad (6)$$

that minimizes total distortion for the whole interval $1 \leq t \leq T$ is derived efficiently by utilizing the DP method (Shiraki, 2004). That is, we have

$$D(t_k) = \min_{t_{k-1} \in R_{k-1}} \left(D(t_{k-1}) + \sum_{t=t_{k-1}}^{t_k} d(y(t), \hat{y}(t)) \right) \quad (7)$$

where

$$R_{k-1} = \{t | t_{k-1} - \delta \leq t \leq t_{k-1} + \delta\}, \quad (8)$$

$D(t_k)$ is an accumulated minimal distortion at t_k and δ is a search range.

In previous work (Hiroya, 2010), we assumed that event timing corresponds to the central point of each phoneme on the time axis and found that the mean errors between the measured EMA-based mid-sagittal articulatory parameters $y(t)$ and the estimated ones $\hat{y}(t)$ were 0.16 mm. This indicates that articulatory motion during speech production can be represented by articulatory positions at the central point of each phoneme and by adequate interpolation functions: articulatory parameters are suitable for a sparse representation. This is because the articulatory parameters for a given phoneme are prominent and because the temporal patterns of the articulatory parameters are simple and smooth. Thus, we had the idea that we would measure vocal-tract MRI images at key frames using sparse sampling. In the next section, we will show the procedure for the proposed method.

3. Procedure for proposed method

We measured the vertical and horizontal positions of articulators, such as the lips, four tongue positions, and lower incisor, using a two-dimensional EMA system (Carstens

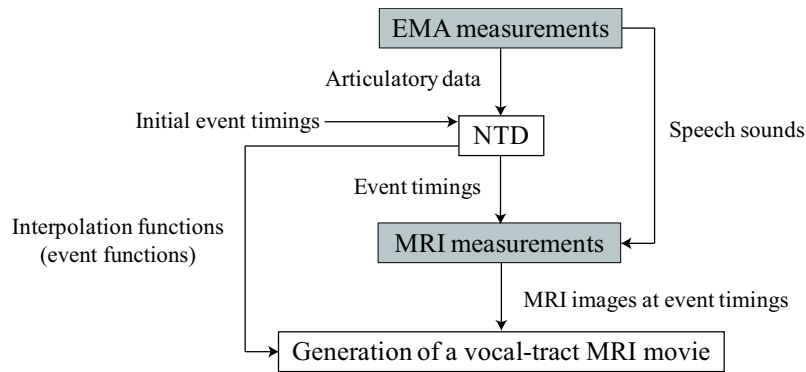


Figure 1. Procedure for proposed method.

AG100). The system can measure the articulators at a sampling rate of 250 Hz and simultaneously record the speech sounds at a sampling rate of 48 kHz.

In NTD, we first set the event vectors as $a_{i,k} = y_i(t_k)$ for the initial event timing t_k , which was labeled at the central point of each phoneme on the time axis manually. Then, the event functions and the event timings are determined by DP and the NMF update rules of Eqs. (4)-(5). Note that event vectors were determined as $a_{i,k} = y_i(t_k)$ depending on the event timings.

Next, we conducted MRI scanning [Siemens MAGNETOM Verio (3T)] for vocal-tract motion during speech production. For the MRI scanning, it was necessary that the same subject in the EMA experiments produce the same sentences as in the EMA experiment through MRI-compatible headphones in order to match the articulatory timing. The MRI scan was conducted at event timing t_k and measured the mid-sagittal section of the vocal tract.

Using the event functions $\phi_k(t)$ obtained by EMA and vocal-tract MRI images (as event vectors) $b_{i,j,k}$ ($1 \leq i, j \leq q$) at event timing t_k , where q is the pixel length, we generated a vocal-tract MRI movie $\hat{x}_{i,j}(t) = b_{i,j,k}\phi_k(t) + b_{i,j,k-1}\phi_{k-1}(t)$ ($t_{k-1} \leq t \leq t_k$) at the sampling rate of 250 Hz. The restriction of $\phi_k(t) = 1 - \phi_{k-1}(t) \in [0, 1]$ would be beneficial for switching the event vectors of EMA with those of MRI. Shape-based interpolation (Grevera and Udupa, 1996) between MRI images was introduced in order to obtain sharp contours of the speech articulators.

4. Results

4.1. Comparison between event functions of MRI and EMA

In the proposed method, we need to use the event functions of EMA to generate a vocal-tract MRI movie. Thus, we should firstly investigate whether the event functions of EMA are similar to those of MRI. We created a vocal-tract MRI movie of the Japanese vowel sequence /aiueo/ based on synchronized sampling with an external trigger and measured EMA-based articulatory parameters of the same sequence. A Japanese male subject produced the vowel sequence to rhythmic repetitions of a noise-burst train composed of a three-beat rhythm for both MRI and EMA experiments. Sampling rates were 67 and 250 Hz for MRI and EMA, respectively. A 256×256 mm² field of view, 256×256 -pixel

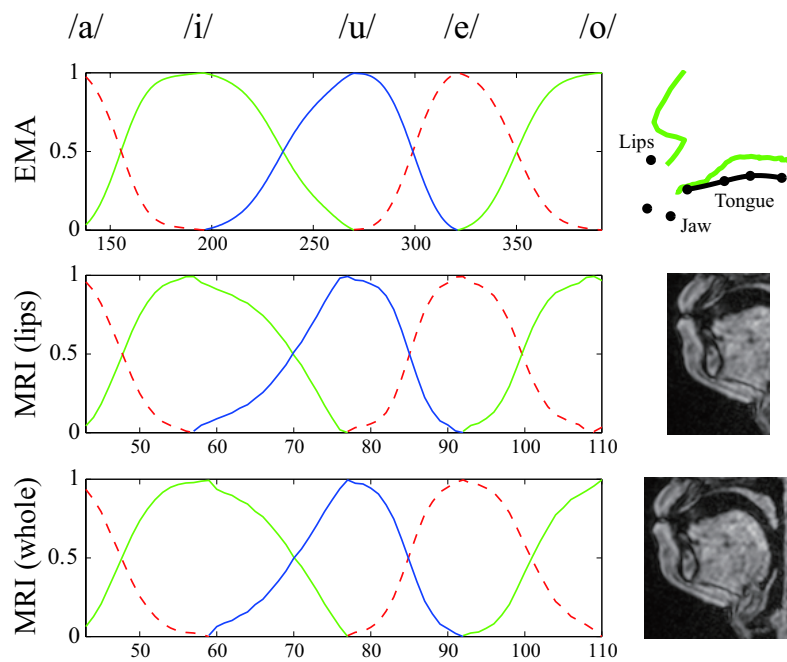


Figure 2. Event functions of EMA, MRI around the lips of the vocal tract and MRI at the whole articulatory area of /aieuo/.

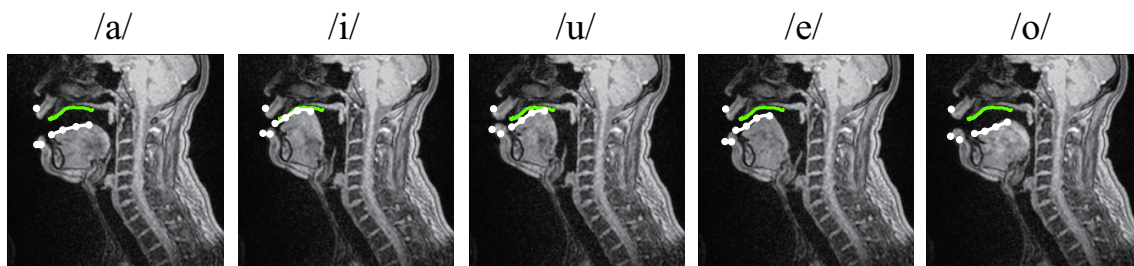


Figure 3. Comparison between upright (EMA) and supine (MRI) postures for /a/ to /o/. White circles represent receiver coil positions of EMA.

image size, and 3-mm slice thickness were used for MRI. All MRI images were converted to 8-bit gray scale. In NTD, the α was set to 10^6 and 10^8 for EMA and MRI, respectively, and δ was around 40 msec. A matrix of the MRI images was converted to a vector for NTD.

Figure 2 shows the event functions of EMA and MRI. Event functions of MRI were calculated for two regions: around the lips of the vocal tract, as compared to EMA, and at the whole articulatory area. Event functions of EMA were similar to those of MRI. Moreover, there is not a large difference between two event functions of MRI. These results indicate that we can use event function of EMA to generate a vocal-tract MRI movie.

The subject spoke in the upright postures for EMA measurements, whereas MRI measurements require supine postures. However, a previous study (Kitamura et al., 2005) suggested individual differences in the effects of gravity on the tongue body between upright and supine postures. As shown in Fig. 3, there is not a large difference in the

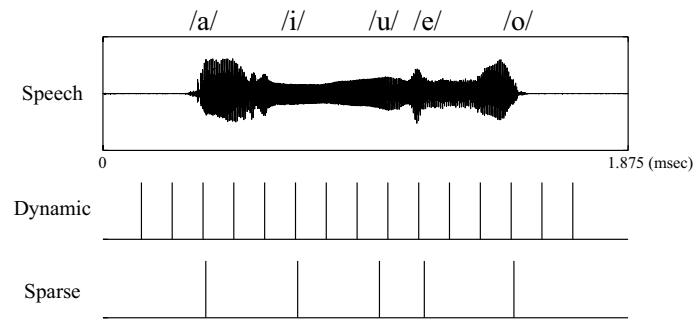


Figure 4. Scan timings for dynamic and sparse MRI of /aiueo/. Vertical bars indicate scan timings.

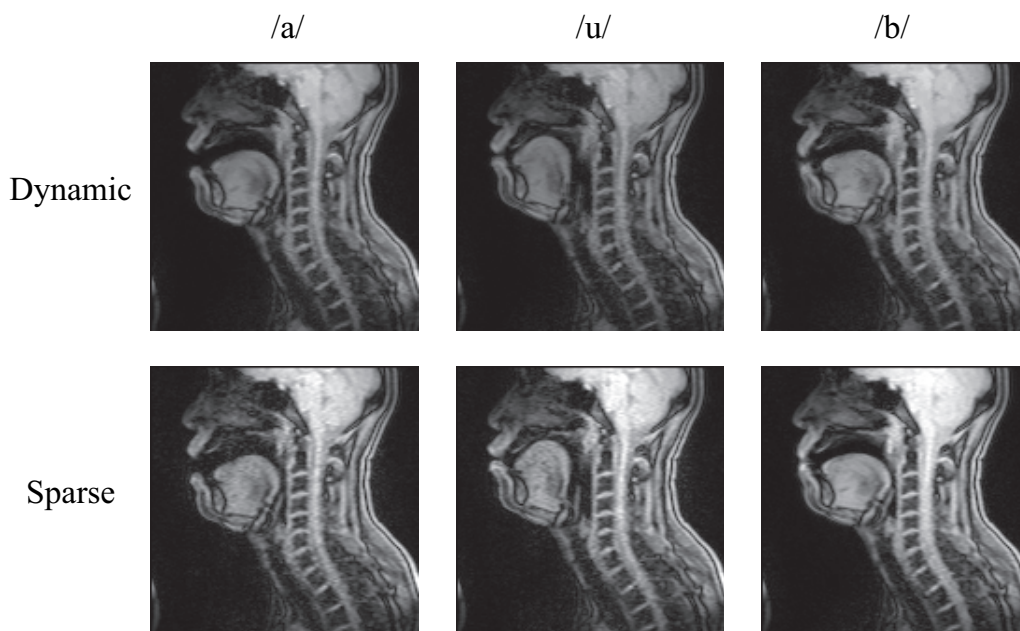


Figure 5. Vocal-tract images of dynamic and sparse methods. /a/ and /u/ of /aiueo/, and /b/ of /tabako/.

vocal-tract information between upright (EMA) and supine (MRI) postures of /a/ to /o/ for the subject.

4.2. Comparison between dynamic and sparse MRI

We compared the proposed method, called sparse MRI, with a high-speed scanning method, called dynamic MRI. Both TRs were 100 msec. A 144×144 -pixel size and 10-mm slice thickness were used for MRI. All MRI images were converted to 8-bit gray scale. Dynamic MRI scanned at the rate of 9 Hz and sparse MRI scanned at event timing t_k in the mid-sagittal section of vocal tract (Fig. 4). For both scanings, the subject tried to speak while he listened to his own voice from the EMA experiment through MRI-compatible headphones in order to match the articulation timing. The Japanese vowel sequence /aiueo/ and the word /tabako/ were used. For /tabako/ of sparse MRI, the number of repetitions required was two because TR was larger than the durations between successive event timings. The image of dynamic MRI at the central point of each phoneme

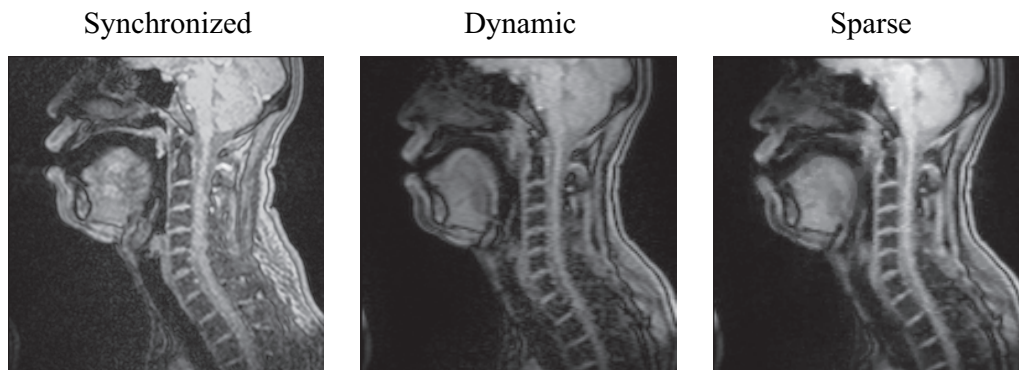


Figure 6. Vocal-tract images between /e/ and /o/ for the synchronized sampling, dynamic, and sparse methods.

was obtained by linear interpolation between two successive images. Figure 5 shows that vocal-tract image at the central point of /a/ is the same for both sparse and dynamic MRI since scan timing of dynamic and sparse MRI was almost the same. However, sparse MRI is better than dynamic MRI for the lip protrusion for /u/ and lip closure for /b/. This indicates that sparse MRI can measure the most important vocal-tract information for each phoneme, compared with dynamic MRI.

4.3. Generation of a vocal-tract MRI movie

Figure 6 shows vocal-tract images for the synchronized sampling, dynamic, and sparse methods in the same timing between /e/ and /o/ of /aiueo/. The image of sparse MRI was obtained with the proposed method and the image of dynamic MRI was obtained by linear interpolation between two successive images. We can see that the image of sparse MRI is more similar to that of synchronized sampling than that of dynamic MRI because the adequate interpolation functions were used for sparse MRI. The quality of the vocal-tract MRI movie obtained by sparse MRI was high compared to that for dynamic MRI.

5. Conclusions

We presented a method for generating a vocal-tract MRI movie based on sparse sampling. This is the first study to generate a vocal-tract MRI movie during speech production using the high temporal resolution of EMA and high spatial resolution of MRI. The method will be useful for constructing a large database of vocal-tract MRI movies and understanding speech production mechanisms.

Acknowledgments

The authors thank Drs. H. Gomi of NTT CS Labs and S. Takano of ATR BAIC for many useful and helpful discussions, and Dr. T. Mochida of NTT CS Labs for helping with the EMA experiments. This study was partly supported by JSPS KAKENHI (21300071).

References

- Atal, B. Efficient coding of LPC parameters by temporal decomposition. In *ICASSP*, pages 81–84, 1983.

- Grevera, G. and Udupa, J. Shape-based interpolation of multidimensional grey-level images. *IEEE Med. Imaging*, 15(6):881–892, 1996.
- Hiroya, S. Non-negative temporal decomposition of speech parameters. In *Proc. ICASSP*, pages 5066–5069, 2010.
- Kaburagi, T., Wakamiya, K., and Honda, M. Three-dimensional electromagnetic articulography: A measurement principle. *J. Acoust. Soc. Am.*, 118(1):428–443, 2005.
- Kim, S.-J. and Oh, Y.-H. Efficient quantisation method for LSF parameters based on restricted temporal decomposition. *Electronics Letters*, 35(12):962–964, 1999.
- Kiritani, S., Itoh, K., and Fujimura, O. Tongue-pellet tracking by a computer-controlled X-ray microbeam system. *J. Acoust. Soc. Am.*, 57(6):1516–1520, 1975.
- Kitamura, T., Takemoto, H., Honda, K., Shimada, Y., Fujimoto, I., Syakudo, Y., Masaki, S., Kuroda, K., Oku-uchi, N., and Senda, M. Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. *Acoust. Sci. & Tech.*, 26(5):465–468, 2005.
- Lee, D. and Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- Masaki, S., Tiede, M., Honda, K., Shimada, Y., Fujimoto, I., Nakamura, Y., and Ni-nomiya, N. MRI-based speech production study using a synchronized sampling method. *J. Acoust. Soc. Jpn. (E)*, 20(5):375–379, 1999.
- Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. An approach to real-time magnetic resonance imaging for speech production. *J. Acoust. Soc. Am.*, 115(4):1771–1776, 2004.
- Perkell, J., Cohen, M., Svirsky, M., Mathies, M., Garabieta, I., and Jackson, M. Electromagnetic midsagittal articulometer system for transducing speech articulatory movements. *J. Acoust. Soc. Am.*, 92(6):3078–3096, 1992.
- Perkell, J. *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Number 53 in Research Monograph. MIT Press, Cambridge, 1969.
- Shiraki, Y. Optimal temporal decomposition for voice morphing preserving Δ cepstrum. *IEICE Trans. Fundamentals*, E87-A(3):577–583, 2004.
- Stone, M., Sonies, B., Shawker, T., Weiss, G., and Nadel, L. Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *J. Phonetics*, 11(3):207–218, 1983.
- Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech and Lang. Process.*, 15(3):1066–1074, 2007.