Establishing some principles of human speech production through two-dimensional computational models

Mauro Nicolao, Roger K. Moore

Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK

m.nicolao@dcs.shef.ac.uk, r.k.moore@dcs.shef.ac.uk

Abstract

Human speech production is often described as an optimisation process, which tends to maximise the effectiveness of the communication process minimising the effort involved in the production.

The aim of this paper is to investigate this highly complex problem with two dimensionally reduced spaces corresponding to different computational models. Since the high-dimensional parameter space which usually describes such a problem is often an issue in the optimal-behaviour computation, two-dimensional models are proposed. The first one analyses the best trajectories visiting the proximity of a set of randomly chosen points. The second one explores the F1-F2 vowel space trying to maximise a set of likelihood functions describing some human production characteristics.

Even though such models need further development, some preliminary correspondences can be observed with some of the elements described in the most popular theories for human speech production. For example, the distance between close competitors directly influences the best trajectory computation and, therefore, the effort needed to achieve the desired tasks. The trajectory planning is also controlled by the *degree of motivation* selected to achieve the desired accuracy: the higher the motivation, the more the target must be addressed.

Index Terms: human speech production model, reactive production model, hyper/hypo-articulation model, optimisation strategies, trajectory planning.

1. Introduction

Human talkers continuously adjust their speech production while they are speaking. One of the first researchers who observed such behaviour was Lombard [1] almost a century ago and, after that study, several other theories have been proposed. Among others, Lindblom's H&H (hypo-hyper) theory [2] affirms that such modifications could be seen as a balancing process in which the talker tries to maximise the success of his communication minimising the effort involved in production.

The hyper/hypo-articulated speech is intrinsically related to the effort that the talker puts in his speech production and it is influenced by his motivation or by the contingencies which may appear in the space. E. g. an utterance can be hyper-articulated because the talker autonomously decides to make his production extremely clear, because the environment forces him to compensate, or because it is imperative to avoid confusion between phonologically similar words.

Principles ruling such an optimisation are not completely established. It might be the result of the continuous speech monitoring by the talker in order to keep it as close as possible to a desired phonetic plan [3]. The control loop could be done assessing the acoustic outcome only, or with the somatosensory system also to achieve prompter reactions to sudden changes [4]. This process is often described as an attempt to satisfy the listener's needs modelled inside the talker's mind by a listener's emulation [5]. The speech modification might be modelled as the result of a previously learned modification of the speech quality (e.g. speech energy reallocation in the time and frequency domain [6]) or, eventually, the product of the enhancement/reduction of the acoustic distance between competing phones in order to minimise possible misunderstandings. In previous experiments [7, 8, 9], it was shown that a data-driven linear transformation which controls such phonetic contrast can be used to tune the degree of synthesised speech intelligibility. These results were found compatible with what most humans do in adverse condition [10].

In order to establish some principles on how humans control speech production, the use of proper computational models might be useful. When parametric representation of speech is given, multidimensional acoustic space is also defined and utterances can be modelled as the parameter vector temporal evolution. If a likelihood function is also defined for every point in such space, speech production turns into an optimisation process which aims to create the trajectory which navigates through the most likely points. It can be assumed that several things influence such function: sequence of targets, trajectory evolution, competing-target density, possible external disturbances, etc.

Though, handling the great number of variables involved in parametric speech representation is a overwhelming problem in the investigation of optimal behaviour. E.g. a standard HMM-based speech synthesiser could have a vector dimension of about 200 elements. In such highly complex spaces, a simple visual representation of the problem, which can be crucial to assess the different strategies, is highly unlikely.

Motivated by these needs, two dimensionally-reduced spaces are introduced to be used as frameworks to test different optimal-trajectory search strategies. The future goal is to extend them to the original multi-dimensional space. Even though they are just simple descriptive models, their observation might, by some extent, establish some fundamental principles still valid in the high dimensional problem. Some minimum assumptions are made in order to guarantee a certain degree of connection to the real problem while complexity is reduced.

- A visual intuitive representation is helpful, therefore a two-dimensional space should be chosen.
- The main goal in this trajectory planning should be to visit a target sequence in a precise order.
- The simplified space should be defined by a set of points to identify targets and by some likelihood functions to describe the area between them.

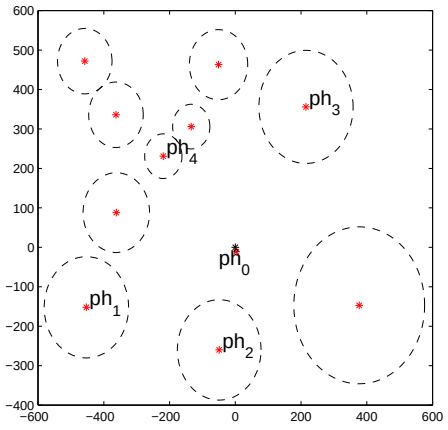


Figure 1: Example of two-dimensional space with 11 random points. The SZs are displayed with dash-lined circles. Four targets ($\mathbf{ph}_{1..4}$) and neutral position, $\mathbf{ph}_0$, are also shown.

- The optimal trajectory might vary at every step as function of current position and of surrounding local space.
- The trajectory-evolution speed should not be fixed, but it should be dependant on the *intensity* of the stimulus: i.e. the farther the trajectory is from the target, the more urgent the movement towards the it should be[11].

In the following sections, two models, which adopt these guidelines, are proposed. Being intuitive and flexible spaces, many different strategies to navigate them can be used and some hints can be extended to real acoustic space.

2. First space: randomly-chosen points

As mentioned, reduction from a high-dimensional space is the most important simplification needed in order to better handle optimal trajectory computation problem. Even if this space have a weak connection to the original acoustic space, it roughly reminds of vowel space. Anyway, in this model more emphasis was put on the navigation techniques of a two-dimensional space rather than on its relationship with the original acoustic space. Hence, the space is defined by a set of randomly-chosen *points*, $\{\mathbf{p}_n\}$, $n = 1..N$. Some of them are named as *targets*, $\{\mathbf{ph}_l\}$, $l = 1..L$, while the $N - L$ inactive points represent obstacles for the trajectory, $\{\boldsymbol{x}_k\}$, $k = 1..T$, to avoid.

Trajectory goal is to *visit* every target in the right order. At each step, one target only is active and $\boldsymbol{x}_k$ needs to go closer to it than to any other surrounding point in order to label current target as visited. Hence, a circular area around each point $\mathbf{p}_n$ is defined. It is named *Safe Zone* for $\mathbf{p}_n$, or SZ($\mathbf{p}_n$) because, when the trajectory is within this area, the related point, $\mathbf{p}_n$, can be *safely* considered as visited. In an ideal vowel space, all the points in that area can be thought as set of *recognisable* realisations of the phone $\mathbf{p}_n$. SZ($\mathbf{p}_n$) radius is different for each point and it is defined as half of the distance between $\mathbf{p}_n$ and the closest among the other points which are also called *competitors*. An example of such space can be found in Figure 1.

The resulting trajectory is influenced by a weighting factor which controls the SZ size. This controlling factor can be interpreted as a measurement of the *motivation* involved in the creation of the trajectory: i.e. how much the system allows for mistakes among competitors. The number of steps needed by the system to complete the path is therefore a direct measurement of this effort.

At the k-th step, one target only, $\mathbf{ph}_l$, can be active. All other positions consequently turn into competitors. The visiting order is also important, therefore a target switching expression is chosen to decide whether to switch the active target:

$$\mathbf{ph}_k = \begin{cases} \mathbf{ph}_0 & k \leq 0 \\ \mathbf{ph}_l & \text{if } \mathbf{ph}_{k-1} = \mathbf{ph}_l \text{ and } \boldsymbol{x}_k \notin SZ_{\mathbf{ph}_l} \\ \mathbf{ph}_{l+1} & \text{if } \mathbf{ph}_{k-1} = \mathbf{ph}_l \text{ and } \boldsymbol{x}_k \in SZ_{\mathbf{ph}_l} \\ \mathbf{ph}_0 & k \geq T \end{cases} \quad (1)$$

where $\mathbf{ph}_0$ represents a neutral position and l is the target index.

Two different strategies were tested to update the trajectory in this space. The first and simplest one aims to find the shortest path which visits each target. The second one forces the trajectory also to avoid competing SZs.

2.1. Minimising the distance to target

Minimising the distance between $\boldsymbol{x}_k$ and target SZ is the first criterion used to compute the desired trajectory.

The resulting trajectory, $\{\boldsymbol{x}_k\}$, is composed by straight lines connecting the target SZs. Its evolution is described by the following equation

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + v(d_k) \cdot \Delta\boldsymbol{x}_k^{\text{ph}} \quad (2)$$

where $\Delta\boldsymbol{x}_k^{\text{ph}} = (\mathbf{ph}_k - \boldsymbol{x}_k)$ is the vector identifying the trajectory direction. The trajectory speed, $v(.)$, is function of the distance to target, $d_k = \|\text{ph}_k - \boldsymbol{x}_k\|_2$. This relationship takes inspiration from Piron's law [12], which states that mean human response time to stimulus are quicker when this is stronger. The law can be expressed by an exponential function. Considering the distance to target can be the stimulus, the speed function can be expressed by

$$v(d_k) = a_1 \cdot e^{\left(d_k + \frac{b_1 \cdot b_2}{b_3}\right)^2} \quad (3)$$

where a_1, b_1, b_2, and b_3 are empirical constants. Example of resulting trajectories are displayed in Figure 2.

In Figure 2a, Figure 2b and Figure 2c, the SZs have different sizes depending on the applied scaling factors. As already mentioned, this value is related to the effort involved in the trajectory creation. High motivation implies that the system is *making the effort* to be more accurate in visiting targets and this influences the correspondent SZ sizes.

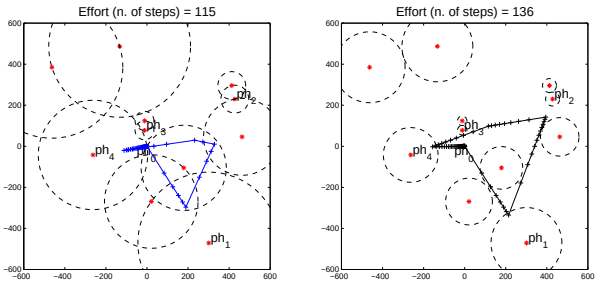
This strategy represents a trajectory-computing algorithm controlled by an important factor in speech production such as motivation, and it has clearly issues related to its simplicity. E.g. nothing prevents $\boldsymbol{x}_k$ to go inside other-point SZs while it is moving towards the target. This means that some points can be marked as visited even though they are not targets and the system creates a different path from the desired one.

2.2. Maximising the distance from competitors

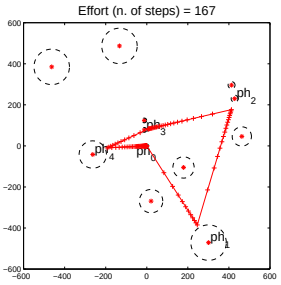
Since the previous strategy often creates trajectories which go into competitor SZs before ending in the right target one, further constraints are needed.

A correction factor, $\boldsymbol{A}(\theta_k)$, is therefore applied to the direction vector in (2). The equation describing the evolution of the model becomes

$$\boldsymbol{x}_{k+1} = \begin{cases} \boldsymbol{x}_k + v(d_k)\Delta\boldsymbol{x}_k^{\text{ph}} & \text{if } \boldsymbol{x}_{k+1} \not\subset \text{SZ}_{ph_{j\neq k}} \\ \boldsymbol{x}_k + v(d_k)\boldsymbol{A}(\theta_k)\Delta\boldsymbol{x}_k^{\text{ph}} & \text{if } \boldsymbol{x}_{k+1} \subset \text{SZ}_{ph_{j\neq k}} \end{cases} \quad (4)$$

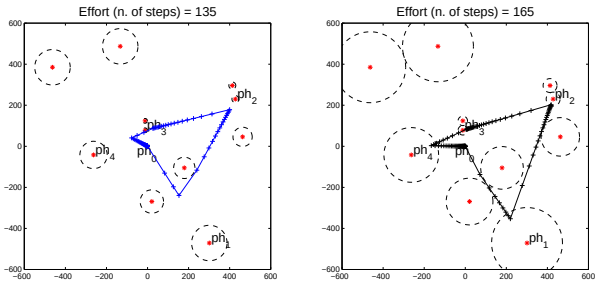


(a) Low motivation (b) Medium motivation

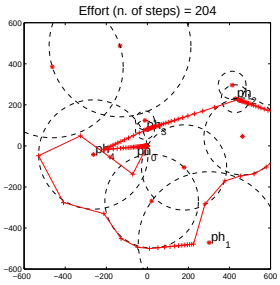


(c) High motivation

Figure 2: Trajectories with different degrees of motivation to reach the same SZ($\mathbf{ph}_{1..4}$) of Figure 1 through the shortest path. Note that SZs are here scaled according motivation.



(a) Low motivation (b) Medium motivation



(c) High motivation

Figure 3: Trajectory with different degrees of motivation to reach the same SZ($\mathbf{ph}_{1..4}$) of Figure 1 by maximising the distance from competitors. Note that circles here show the minimum distance from competitors that should be guaranteed.

where $\Delta\boldsymbol{x}_k^{\text{ph}}$ and $v(d_k)$ are as per (2) and (3). $\boldsymbol{A}(\theta_k)$ is the matrix representing the a θ_k-degree rotation, $\theta_k \in [0, 2\pi]$. Examples of such trajectories are displayed in Figure 3.

In this case, the motivation factor controls the target SZ sizes as in par. 2.1 and the competitor SZ sizes with the inverse of its value. This means that the higher the motivation is, the smaller the target SZ is and the bigger the competitor SZs to avoid are (see Figure 3c).

This strategy is more sophisticated than the previous one and it shows the importance of controlling the distance from competitors in order to minimise false target recognition.

Nonetheless, it still has limitations. One of these stands in the binary decision function deciding whether the target $\mathbf{ph}_l$ was visited.

3. Second space: the vowel space

The previous model is clearly a quite drastic simplification which has weak relationship to some acoustic representation of speech. A further step towards the real problem is hence introduced. This evolution is inspired by the affinity between the previous space and one of the most common vowel parametrical space, the F1-F2 chart.

The points $\{\mathbf{p}_n\}$ which define such two-dimensional space are chosen to be the mean F1-F2 values extracted from the vowels in the CMU-arctic SLT corpus (American English female voice). These values, $\boldsymbol{\mu}_{\mathbf{p}_n}$, along with relative variances, $\boldsymbol{\sigma}_{\mathbf{p}_n}$, are used to define some Gaussian mixture functions, which represent a statistical description of the likelihood of being close to a vowel-formant mean value, see Figure 4. This function substitutes the previous SZ-based criterion.

The $\{\mathbf{ph}_l\}$ targets are the subset of vowels to be pronounced.

3.1. Optimal trajectory computation

Inspired by some state-of-the-art strategies adopted for the trajectory planning in physical environments [13], a combination of different target functions can be considered to compute the optimal trajectory.

Given a set of target vowels, $\mathbf{ph} = \{\mathbf{ph}_l\}$, the optimal trajectory, $\boldsymbol{x}'$, results from maximising the following equation

$$\boldsymbol{x}' = \arg\max_{\boldsymbol{x}} G(\boldsymbol{x}, \mathbf{ph}) \tag{5}$$

with $G(\boldsymbol{x}, \mathbf{ph})$ which depends on the current position and the target sequence. It can be expressed by the following sum of weighted functions

$$G(\boldsymbol{x}, \mathbf{ph}) = c_1 \cdot G_1(\boldsymbol{x}, \mathbf{ph}) + c_2 \cdot G_2(\boldsymbol{x}) + c_3 \cdot G_3(\boldsymbol{x}, \mathbf{ph}) \tag{6}$$

where $\{c_j\} \geqslant 0$ are the parameters to control the motivation associated to the system, and the $\{G_j\}$ functions are specified as in the following paragraphs.

A target is assumed to be visited when current trajectory $\boldsymbol{x}_k$ reach a position which maximises $G(\boldsymbol{x}, \mathbf{ph})$ for the related target. Although (5) and (6) are quite simple, they are very effective to control the behaviour of this formant generator.

Evolution of the target sequence is also fundamentals. Extending the principles of (1) in this space, $\mathbf{ph}_k$ can be expressed

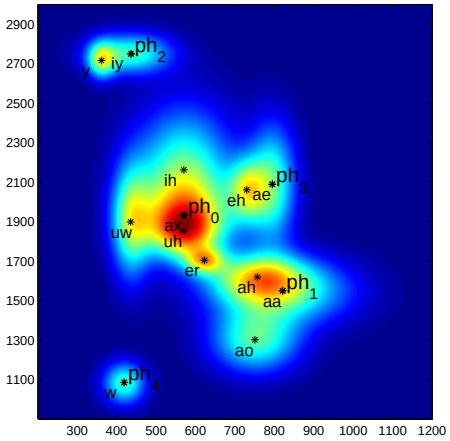


Figure 4: Example of F1-F2 vowel space representing 11 English vowels. All of them are plotted to have same likelihood. Four targets ($\mathbf{ph}_{1..4}$) along with the neutral position $\mathbf{ph}_0$ are also shown. Red areas have the higher likelihood. Phone labels are displayed with the 'CMU Pronouncing Phoneme Set' (*http://www.speech.cs.cmu.edu/cgi-bin/cmudict*)

as

$$\mathbf{ph}_k = \begin{cases} \mathbf{ph}_0 & k \leq 0 \\ \mathbf{ph}_l & \text{if } \mathbf{ph}_{k-1} = \mathbf{ph}_l \text{ and } \Delta G_k \geq \epsilon \\ \mathbf{ph}_{l+1} & \text{if } \mathbf{ph}_{k-1} = \mathbf{ph}_l \text{ and } \Delta G_k < \epsilon \\ \mathbf{ph}_0 & k \geq T \end{cases} \quad (7)$$

where $\Delta G_k = \|G(\boldsymbol{x}_k, \mathbf{ph}_k) - G(\boldsymbol{x}_{k-1}, \mathbf{ph}_{k-1})\|$ and ϵ is a threshold value.

3.1.1. *First function*

The first term in (6) is a Gaussian function which aims to describe the likelihood to be close to the $\mathbf{ph}_k$ target-phone.

$$G_1(\boldsymbol{x}_k, \mathbf{ph}_k) = \frac{1}{2\pi\boldsymbol{\sigma}_{\mathbf{ph}_k}} e^{\left(-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu}_{\mathbf{ph}_k})^\top \boldsymbol{\sigma}_{\mathbf{ph}_k}^{-1} (\boldsymbol{x}_k - \boldsymbol{\mu}_{\mathbf{ph}_k})\right)} \quad (8)$$

where

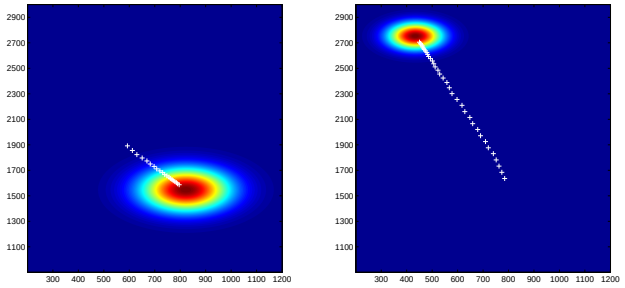
$$\boldsymbol{\mu}_{\mathbf{ph}_k} = \begin{bmatrix} \mu_{\mathbf{ph}_k}^{F1} & \mu_{\mathbf{ph}_k}^{F2} \end{bmatrix}^\top \text{ and } \boldsymbol{\sigma}_{\mathbf{ph}_k} = \begin{bmatrix} \sigma_{\mathbf{ph}_k}^{F1} & 0 \\ 0 & \sigma_{\mathbf{ph}_k}^{F2} \end{bmatrix}$$

are respectively the 2x1 mean vector and the 2x2 diagonal covariance matrix of the first and second formant distribution. It is assumed to have no correlation between the two formant values.

In Figure 6, it is shown that, without any limitation given by G_2 or G_3 (i.e. $c_2 = c_3 = 0$), the trajectory almost reach the mean positions of every phone. In this case, the trajectory can be assumed to have achieved a completely realised (i.e. fully-articulated) version of the target vowel sequence.

3.1.2. *Second function*

The second term, $G_2(\boldsymbol{x}_k)$, in the optimisation function depends on the current position only and it is motivated by the hypothesis that an unique Low-Energy (LE) attractor exits for vowels in human speech production. This attractor can be identified in



(a) $\mathbf{ph}_k$ label = 'aa' (b) $\mathbf{ph}_k$ label = 'iy'

Figure 5: Examples of trajectories to visit different targets when G_1 alone is activated in G ($c_1 = 1$, $c_2 = 0$, $c_3 = 0$).

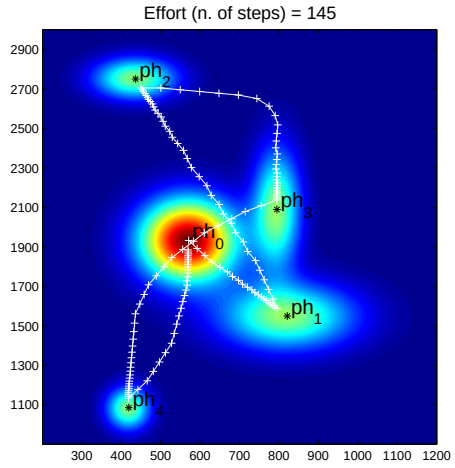


Figure 6: Complete resulting trajectory in the two-dimension vowel space when G_1 alone is activated.

the mid-central vowel *schwa* [7]. This LE-emulation function describes the likelihood of being close to the central vowel and it can be expressed by the following single Gaussian function:

$$G_2(\boldsymbol{x}_k) = \frac{1}{2\pi\boldsymbol{\sigma}_{LE}} e^{\left(-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu}_{LE})^T \boldsymbol{\sigma}_{LE}^{-1} (\boldsymbol{x}_k - \boldsymbol{\mu}_{LE})\right)} \quad (9)$$

with $\boldsymbol{\mu}_{LE} = \begin{bmatrix} \mu_{LE}^{F1} & \mu_{LE}^{F2} \end{bmatrix}^\top$ and $\boldsymbol{\sigma}_{LE} = \begin{bmatrix} \sigma_{LE}^{F1} & 0 \\ 0 & \sigma_{LE}^{F2} \end{bmatrix}$. $\boldsymbol{\mu}_{LE}$ is assumed to have the same values as $\boldsymbol{\mu}_{\text{schwa}}$ while the variance $\boldsymbol{\sigma}_{LE}$ must have a big enough value to allow for every realisation in such a space.

In Figure 8, it can be seen that, as expected, the trajectory never escape from the attraction of the LE point.

3.1.3. *Third function*

The third term in $G(\boldsymbol{x}, \mathbf{ph})$ models the likelihood for a point in the space to be a not-target phone. It can be expressed as the following Gaussian mixture function:

$$G_3(\boldsymbol{x}_k, \mathbf{ph}_k) = \sum_{i=1, \mathbf{ph}_i \neq \mathbf{ph}_k}^{N} \frac{1}{2\pi\boldsymbol{\sigma}_{\mathbf{ph}_i}} e^{\left(-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu}_{\mathbf{ph}_i})^\top \boldsymbol{\sigma}_{\mathbf{ph}_i}^{-1} (\boldsymbol{x}_k - \boldsymbol{\mu}_{\mathbf{ph}_i})\right)} \quad (10)$$

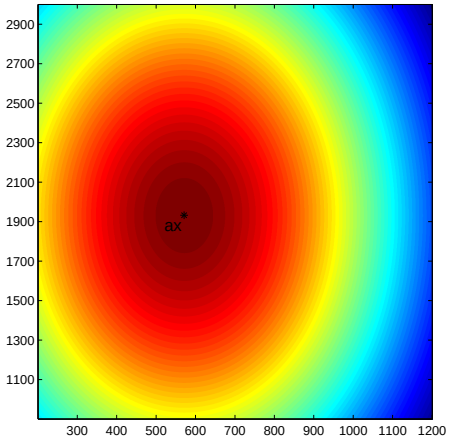


Figure 7: Example of likelihood distribution when G_2 alone is activated ($c_1 = 0$, $c_2 = 1$, $c_3 = 0$). $\sigma_{LE}^{F1} = 500$ and $\sigma_{LE}^{F2} = 1300$.

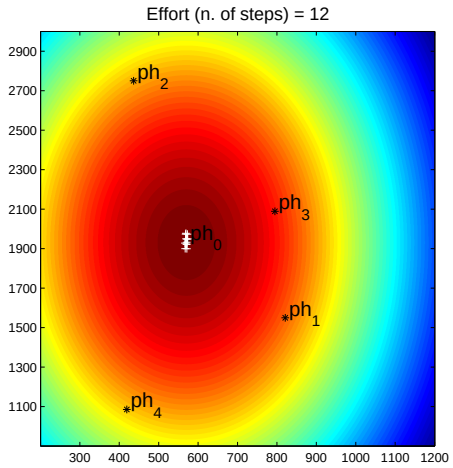


Figure 8: Complete resulting trajectory in the two-dimension vowel space when G_2 alone is activated. $\{\boldsymbol{x}_k\}$ never escape from the LE point because there are no other active attractors.

where N is the number of phones in the acoustic space and $\mathbf{ph}_k$ the current target whose likelihood should not be included in the mixture.

It is worth to notice that the parameter c_3 in (6) must be a negative number because the likelihood of being close to non-target should be minimised in the overall function.

Being the peripheral zone of the vowel space the most likely area far from all competitors (see Figure 9a and Figure 9b), it can be observed that the trajectory resulting from this function can easily escape outside the space boundaries.

3.1.4. Trajectory computation

The optimisation function (6) described in the above paragraphs is used to compute the trajectory at every step, using an heuristic algorithm to reach the most likely point. This is, by definition, a sub-optimal algorithm but it is chosen to understand what degree of prediction is needed in such decision process. In particular, it is important to study whether the optimisation can be done locally or it should be done taking into account of the whole path. In details, the trajectory update expression is

$$\boldsymbol{x}_{k+1} = \arg\max_{\boldsymbol{x}} G(\boldsymbol{C}_k, \mathbf{ph}_k) \tag{11}$$

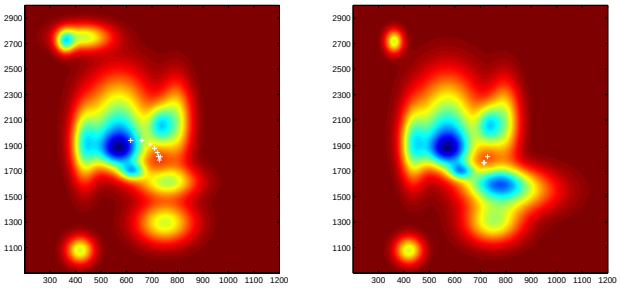
where $\boldsymbol{C}_k$ is a circular point subset around $\boldsymbol{x}_k$. Such a circle is used to reduce the complexity of $G(\boldsymbol{x}, \mathbf{ph})$ computation and its radium, r_k, is dependant on the distance to the target:

$$r_k = r_0 \cdot d_k = 0.1 \cdot \|\mathbf{ph}_k - \boldsymbol{x}_k\| \tag{12}$$

with r_0 is a arbitrary scaling factor.

The search for the global maximum is extremely difficult with this heuristic optimisation because several local maximum appear in such a space. The main consequence is that the trajectory could easily be trapped in some local maximum point which is not the desired target (see Figure 11).

Even though this system still exhibit several limitations, especially in the planning algorithm, some connection with some of the most popular theories of speech production can be observed. The differences among the $\{c_j\}$ parameters can be correspondent to *motivation* which the system is allowed to use



(a) $\mathbf{ph}_k$ label = 'aa' (b) $\mathbf{ph}_k$ label = 'iy'

Figure 9: Example of likelihood distribution when G_3 alone is activated in G ($c_1 = 0$, $c_2 = 0$, $c_3 = -1$). In peripheral area the likelihood of being far from not-target phone is higher.

to generate the trajectory. The LE attractor function, G_2 controls the degree of vowel dispersion/reduction and G_3 represent a way to manipulate phonetic contrast.

In the end, it is worth to emphasise that every point in the trajectory can have a direct correspondence to a voiced sound and, therefore, every trajectory can be easily synthesised with a formant synthesiser such as the Holmes synthesiser [14].

4. Conclusions and further direction

The models proposed in this paper represent a useful framework to study some issues related to some optimisation problems in human speech production. The problem was in fact modelled as a trajectory optimisation task in a multidimensional space to reduce the intrinsic complexity.

The biggest advantage of these models is the dimensional reduction, which allows to represent the events in a two-dimensional space and to visualise the resulting trajectories as they were physical paths among obstacles. Even though these methods are still at an early development stage, it has been

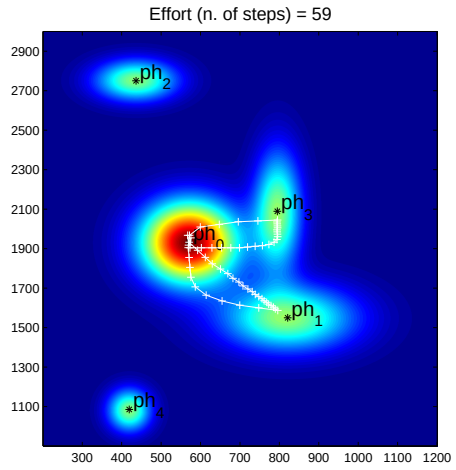


Figure 10: Trajectory in the vowel space when G_1 and G_2 are activated ($c_1 = 10$, $c_2 = 10^{-3}$, and $c_3 = 0$). $\mathbf{ph}_2$ and $\mathbf{ph}_4$ are not achieved because they are too far from the LE attractor.

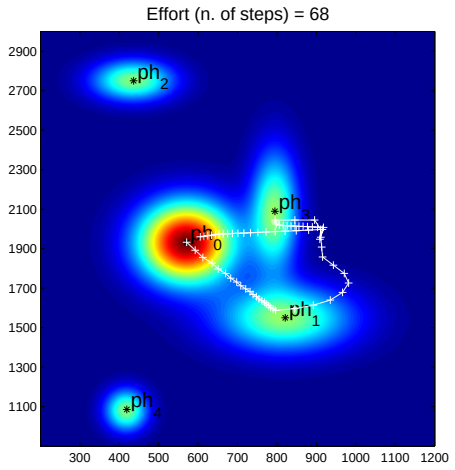


Figure 11: Trajectory in the vowel space when all terms in G are active ($c_1 = 10$, $c_2 = 1 \cdot 10^{-3}$, $c_3 = -1 \cdot 10^{-3}$). It is worth to notice that not all the target point are reached due to some local attraction.

observed that some links with some of the major elements of human speech production can be established. A more careful comparison with similar computational models has indeed to be done in order to evaluate the performances of this model.

Further development is needed to compute the optimal path. Some trajectory-evolution prediction along with some past position memory would improve performance especially in order to avoid local maximums.

Another great opportunity of such models is its flexibility. Potentially, many further functions can be added in order to model different aspect of speech production. For example, a masking disturbance could be described as an obstacle to avoid in such simplified spaces. Hence, the new optimisation would compute the trajectory avoiding the new obstacles.

The development of such controlling functions represents a great opportunity to create a better computational model to be used in automatic speech synthesis as well. This would be extremely important, since most of the state-of-the-art synthesis systems exhibit a rather limited range of speaking styles as well as an inability to react to conditions in which they operate.

5. Acknowledgements

The research leading to these results was funded by the EU-FP7 under grant agreement n. 213850 - SCALE.

6. References

[1] É. Lombard, "Le Signe de l'Elevation de la Voix - The sign of the rise in the voice," *Ann. Maladiers Oreille, Larynx, Nez, Pharynx - Annals of diseases of the ear, larynx, nose and pharynx*, vol. 37, pp. 101–119, 1911.

[2] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," *Speech production and speech modelling*, vol. 55, pp. 403–439, 1990.

[3] W. J. M. Levelt, *Speaking: From intention to articulation*. The MIT press, 1989.

[4] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, no. 2006, pp. 280–301, Jul. 2005.

[5] R. K. Moore, "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1176–1188, Sep. 2007.

[6] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *INTERSPEECH 2010*, Makuhari, Chiba, Japan, Sep. 2010, pp. 1636–1639.

[7] R. K. Moore and M. Nicolao, "Reactive Speech Synthesis: Actively Managing Phonetic Contrast Along an H&H Continuum," in *ICPhS 2011*, Hong Kong, China, Aug. 2011, pp. 1422–1425.

[8] M. Nicolao and R. K. Moore, "Consonant production control in a computational model of hyper & hypo theory (C2H)," in *LISTA workshop 2012*, Edinburgh, UK, May 2012.

[9] M. Nicolao, J. Latorre, and R. K. Moore, "C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech," in *INTERSPEECH 2012*, Portland, OR, Sep. 2012, pp. 1–4.

[10] D. R. van Bergem, "Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress," *Speech Communication*, vol. 16, pp. 329–358, Jan. 1995.

[11] L. van Maanen, R. P. P. P. Grasman, B. U. Forstmann, and E.-J. Wagenmakers, "Piéron's law and optimal behavior in perceptual decision-making," *Frontiers in Neuroscience*, vol. 5, pp. 1–15, Dec. 2011.

[12] H. Piéron, "Recherches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices," *L'année psychologique*, vol. 20, no. 20, pp. 17–96, 1913.

[13] M. Svenstrup, T. Bak, and H. J. Andersen, "Trajectory planning for robots in dynamic human environments," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 4293–4298.

[14] J. N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE transaction on Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, Jun. 1973.