# Transcription of the Workshop

# SPEECH PRODUCTION MODELS AND EMPIRICAL EVIDENCE FROM TYPICAL AND PATHOLOGICAL SPEECH

13 May 2024, Grenoble, France

## Outline

ISSUE 2
What problems arise from the fact
language production models ha
developed on the basis of health
(and not pathological spee

*How to cite this document:* Fougeron C., Goldstein L., Guenther F., Lœvenbruck H., Mefferd A., Mücke D., Niziolek C., Parrell B., Perrier P., Ziegler W., Laganaro M. (unpublished manuscript) Transcription of the workshop Speech Production Models and Empirical Evidence from Typical and Pathological Speech. 13 May 2024, Grenoble, France. https://doi.org/10.26037/yareta:cvlb5qujzzc3ti3pgxq62y76ui

# Premise

The Workshop "Speech production models and empirical evidence from typical and pathological speech" has been organized in the context of the **ChaSpeePro** project, which is a collaborative interdisciplinary research project financed by the Swiss National Science Foundation (CRSII5_202228). The project investigates motor speech encoding processes in neurotypical speakers and in speakers with different forms of motor speech disorders with a multidisciplinary approach. The workshop was held in Grenoble on the 13th of May 2024, thanks to the collaboration with Gipsa-Lab. It was aimed at debating in a convivial and constructive atmosphere theoretical positions on speech production and **empirical evidence from both typical and pathological speech** on three major questions:

1. Planning/programming/execution or phonological/phonetic/motor encoding (or other encoding/computing distinctions): how to define the different processes in (motor) speech production?
2. Encoding units/representations in speech production models: which ones, how many different units, how are they selected and combined in larger units?
3. How are different speech modes (whispered, loud, fast, clear, …) encoded/parametrized for production?

In the morning, four **invited talks** addressed and introduced these questions. The four talks were :

- Neural processing stages in speech production, by Frank Guenther (Boston University)
- Sensorimotor adaptation and planning units, by Ben Parrell (University of Wisconsin-Madison)
- Speech modulations and articulatory control mechanisms, by Antje Mefferd (Vanderbilt University Medical Center)
- Issues and challenges from the ChaSpeePro project, by Marina Laganaro and Cécile Fougeron for the ChaSpeePro team (University of Geneva and CNRS/Université Sorbonne-Nouvelle)

In the afternoon four participative round tables debated specific questions:

1. *Speech production processes and units*, with Pascal Perrier (moderator), Louis Goldstein, Frank Guenther, Ben Parrell, Caroline Niziolek
2. *Neurotypical versus impaired speech*, with Doris Mücke (moderator), Louis Goldstein, Frank Guenther, Antje Mefferd, Caroline Niziolek, Pascal Perrier, Wolfram Ziegler
3. *Short-term adaptations and speech modes/styles*, with Hélène Lœvenbruck (moderator), Frank Guenther, Doris Mücke, Ben Parrell, Pascal Perrier
4. *Learning, changing, adapting speech*, with Wolfram Ziegler (moderator), Louis Goldstein, Frank Guenther, Hélène Lœvenbruck, Pascal Perrier

The discussions at the round tables have been recorded with authorization of all speakers. The present document is an edited version of the automatic transcription of the discussions during these four round-tables, each preceded by a summary of the discussions.

## Acknowledgments

We thank for their help in the edition of this document: Julia Chauvet, Monica Lancheros and Marion Bourqui.

# Round table 1: Speech production processes and units

*Pascal Perrier (moderator), Louis Goldstein, Frank Guenther, Ben Parrell, Caroline Niziolek*

## *Summary*

*The main topics addressed in this round table are: the distinction between planning and programming processes, between stored and computed units and the size of the units involved as well as processes and modalities involved in feedback*

***Planning and programming** Planning was described as going from an abstract linguistic representation to a time-varying sensorimotor trajectory, as well as assigning sensory goals. Programming would be adjusting the plan to the specific physical and utterance context; in other words, implementing an optimized controller shaped by contextual constraints. Some emphasized a clear division: planning specifies a trajectory, programming applies a cost function, and control adjusts movement to meet goals. Others argued for more interactive or overlapping processes as programming and execution continuously interact. A consensual distinction was proposed: 'Planning One' = abstract planning and assigning sensory goals, 'Planning Two' = planning in context (i.e., programming).*

***Stored versus computed** The discussants also addressed the questions about what is stored versus computed on-line. There was a shared view that the term "motor program" (or plan) refers to something that can be learned and optimized over. The motor plan/program is shaped by prior experience and will be adapted to the immediate communicative context. It was also mentioned that we do not simply have unlearned patterns on the one hand, and completely learned ones on the other. Rather, what is learned falls somewhere in between because the brain is malleable. Motor speech sequences (e.g., syllables, words) are constructed and refined over time. Novel productions may initially be phoneme-by-phoneme, but with repetition and practice, units become chunked into larger motor programs. In other words, sequences are built and refined over time via optimization. Frequent sequences are more efficient due to repeated optimization.*

***Feedback and perturbation** Across the different contributions, it was mentioned that feedback operates across multiple levels and modalities, and is state-dependent. At lower levels, feedback enables real-time adjustments to variables such as formants, while at higher levels, it supports the monitoring of phonemic errors. In terms of modality, auditory feedback is closely tied to communicative goals and outcomes and its monitoring may be prioritized due to its relevance for intelligibility. In contrast, somatosensory feedback is more directly linked to the physical state of the articulators. However, there was no consensus on whether somatosensory feedback holds equal importance to auditory feedback. For example, it was noted that somatosensation can even support vowel categorization. Another observation is that when auditory feedback is impaired (e.g., noisy environments), reliance on somatosensory feedback increases. It was concluded that both auditory and somatosensory feedback are integrated and valuable. In terms of perturbation, it was thought that there are quick adjustments made on the fly for unexpected perturbations, and some longer-term adaptations tied to the state of the system, for example, being tired or dehydrated—which can also influence the acoustic realization of speech.*

***Size of units** There was general consensus that speech planning is flexible in terms of the units involved—ranging from sub-syllabic elements to full phrases. Rather than committing to a single dominant unit, participants suggested that multiple units may be encoded simultaneously or become relevant under different task demands, contexts, or neural substrates. Evidence from computational models and neuroimaging supports the idea that representations of various sizes can co-exist and may map to distinct brain areas.*

Pascal Perrier (moderator/discussant)

Two main questions will be addressed in this round table:

(a) What is the difference between motor planning and linguistic planning?

(b) Which are the units that are planned, that are controlled?

This morning, at least in the presentations by Frank, Ben, Marina and Cécile[1], we saw that there is no consensus about what is meant by *planning* and by *programming*. Another notion was not explicitly mentioned, namely execution. To my understanding, Ben, Frank, and I would likely agree that there is a stage preceding the motor level—the motor plan—that relates to linguistics.

As for units, my understanding is that while we are uncertain about their exact nature or the linguistic structures involved [in motor speech planning/programming], we do know that there is some sub-syllabic unit that can be grouped together to form larger ones.

## Planning, programming, execution

I will now focus on discussing the second part, which was addressed by Frank and Ben and is more closely related to models.  It is not clear to me, based on Frank and Ben's talks, how planning and programming are defined.

In order to introduce the discussion concisely, I would like to illustrate the evolution of GEPPETO in relation to the notion of planning. GEPPETO is a model that we developed in Grenoble; there have been several versions over the years. One of the versions, developed by Jean-François Patri with the help of Julien Diard and myself, is based on what is known as Bayesian planning[2].

In this model, illustrated in Figure 1, which we consider a model of planning, the process unfolds as follows. We start at the linguistic level, let's say with a sequence of phonemes. These phonemes are represented by ϕ. These are the linguistic content that we aim to achieve. Motor planning begins at this point because it involves specifying goals—specifically, sensory goals in the auditory (A) or somatosensory (S) domains. Additionally, to take into account aspects raised this morning by Marina and Cécile, such as loudness and clarity, the model incorporates a force parameter (W). Thus, phonemes are associated with goals in the physical domain as the first step of planning.

Then, thanks to internal models, we can decide which motor commands are best suited for producing the three phoneme correlates (A, S, W). Since GEPPETO is a Bayesian model, a sequence is represented as a probability distribution over the command space. In this way, we know which is the best probability for producing /ʁ/, /ɛ/, or /a/. And then we have a sequence of goals. At this level, the sequence of goals corresponds to an initial grouping—possibly a syllable, a VCV structure, or a sequence of syllables.

We then optimize this sequence of goals by identifying the best probability associating these three levels, leveraging internal models and the specification of the goals associated to find it.

[1] This refers to the morning presentations, see program here: https://www.unige.ch/fapse/mospeedi/Workshop2024

[2] Patri, J.F., Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. Biological Cybernetics, 109 (6), 611–626

Patri, J.F., Diard, J., & Perrier, P. (2019). Modeling sensory preference in speech motor planning: a Bayesian modeling framework. Frontiers in Psychology, 10, Paper 2339

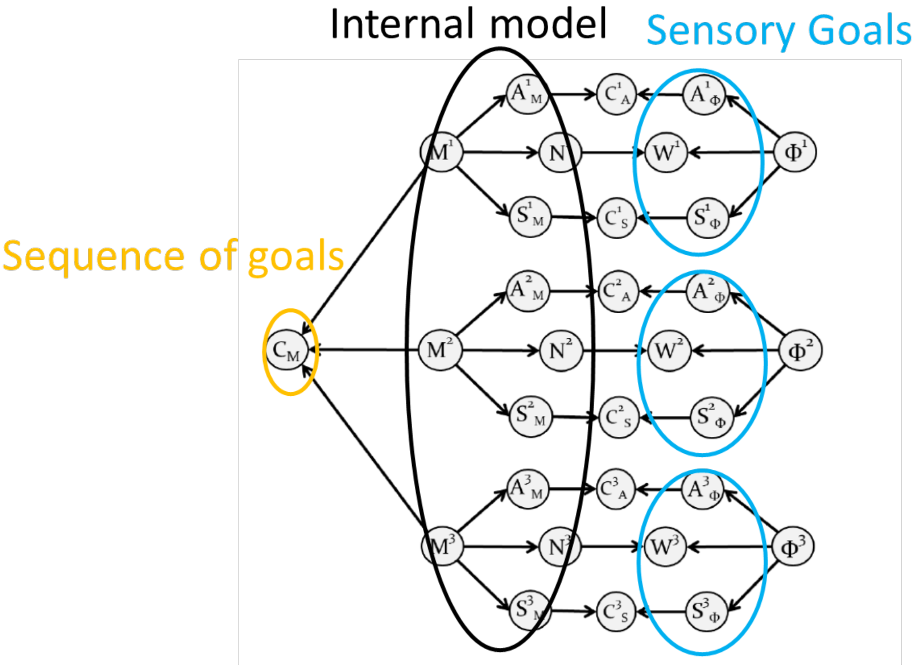

*Figure 1. Motor planning as modeled in Bayesian GEPPETO (Patri et al., 2019). $M^i$: motor commands used for phoneme $F_i$; $A^i{}_M$, $S^i{}_M$ and $N^i$ auditory output, somatosensory output and force level predicted with the internal model for phoneme $F_i$; $A^i{}_F$, $S^i{}_F$ and $W^i$, desired auditory, somatosensory and force level goals associated with phoneme $F_{i.}$*

In our framework, this stage is what we considered ‘planning’. However, we realized that there is a problem with planning as such: if execution were to begin immediately based on this planning alone, there would be no guarantee of reaching the target accurately and at the right time. We then recognized the need to integrate ongoing feedback. This means that within planning, or programming, a certain degree of feedback integration is necessary to provide information about the actual execution of movement. This approach was incorporated into the next version of GEPPETO, called *GEPPETO Optimal Control* (GEPPETO OC). In this model, at the left-hand side of Figure 1, still represents the definition of goals associated with phonemes in terms of acoustic, tactile, and somatosensory goals. The system then determines the appropriate motor command, denoted here as Lambda (λ). Interestingly, these commands are set up as sent to the biomechanical model. By biomechanical model, we mean that what is actually executed does not exactly match what is planned, due to physics interfering with the command. Then, a comparison is made between the actual feedback received and the internal representation of the biomechanical model—what would be the ideal output of the motor command. An optimal estimator then assesses the system’s true state and makes necessary corrections. Importantly, this optimal controller optimizes the cost function and includes effort minimization (a typical aspect). I won’t go into detail on this. So we retain key elements from the first GEPPETO version, including the definition of goals and the concept of a sequence of goals—now with the addition of timing.

In my view, in this model, planning consists of taking into account the goals, the intrinsic timing—when we want the goals to be reached—and the sequence. This implies that programming (or perhaps planning; I’m not sure which term fits best) will be different, and the outcome will be different, depending on the sequence being planned. Planning a syllable will yield a certain result, a VCV sequence another, and a sequence of syllables yet another. And of course, optimization occurs along this path.

What is important here is that it is difficult for me to distinguish between planning, programming, and execution, as all this occurs during execution. In this model, reprogramming is continuous. Of course, if the initial planning works—based on Bayesian GEPPETO, a set of commands for the goal—little adjustment is needed. But if it fails, if effort minimization is not achieved during the initial programming, then planning will be changed. This brings us to the issue: distinguishing between planning, programming, and execution. I understand that, for some of you, programming refers to how commands are implemented in the physical domain. Execution, in turn, involves not only the continuous control of motor output—functioning like a maintenance system—but also the system's response to those commands, which is shaped by biomechanics. While biomechanics can be controlled to some extent, it still imposes inherent constraints.

Frank Guenther

Can I ask you a question, Pascal? How fast does that learning process occur? Do you think that each time you say something new, the entire estimation process takes place?

Pascal Perrier

I don't think so. If this planning, programming, and execution is done often, as you propose in frequent syllables, or in frequent words, it's likely stored. However, in my opinion, what is not explicitly addressed in GODIVA is the process involved when producing a sequence for the first time or when generating a completely new sequence of phonemes. We propose that this process sets up the command and re-actualizes it depending on changes in our physical, auditory, or somatosensory system.

Also, the context or environment in which we speak—since sometimes we cannot hear ourselves, we rely more on somatosensation—may lead to differences.

So, this is not systematically the only process, but it is always possible to engage it if the stored motor program, or pattern, is unlikely to be used.

Frank Guenther

I largely agree. I think one important point to keep in mind is that we do not simply have unlearned things on the one hand, and completely learned things on the other. Rather, everything falls somewhere in between. And so it is artificial in a way to talk about syllable units, etc. Those things are malleable, that is how the brain is. At the same time, however, to make sense of it, I think it's important that we start to break these things down.

Ultimately, I think we have to look at the brain circuits to determine what these things are: are there stored motor programs, and if so, where? We can also collect data from neurons that directly address the questions we are discussing today. In the end, I think that's where we will find answers about the best way to describe these things. It will come from an increasingly better understanding of what is actually happening in the brain.

But specifically regarding GODIVA[3] (see Figure 4) learning new things—when it starts out, it simply learns sequences without regard for what the units are. These units could be phonemes or words; GODIVA just learns to concatenate a sequence that it produces repeatedly into a motor program. But like I said, everything likely falls in between. If it's a word you say frequently, you probably have a rough motor program stored, though much of it still has to be generated online. Ultimately, these are all continuous processes, and in a way, it is artificial to define discrete units. And when we look at a brain diagram, every part of the brain and cortex is interconnected—it's not a simple, bold block diagram. Any of the diagrams we've shown today will only capture part of this complexity, so we inevitably have to break it down in some way.

Pascal Perrier

Let’s return to the question of planning and programming. If we consider your latest results, showing that depending on whether the structure is a usable (frequent) structure or not, you have approximately the same reaction time, but then it goes more slowly than what you showed. So, interpreting this, in the case of a frequent sequence of sounds, we could rely on a syllabary of stored words. And the process should be faster than if we do not have to do it.

Then, what Marina observed indicates an apparent contradiction: initially, the process does not take more time, yet later, it does. This could support the idea that programming and execution—or planning and execution—occur in parallel. In this view, you plan and execute simultaneously, but with a longer processing time when stored patterns are unavailable. The lack of reaction time differences before speech onset suggests that the initial phase of processing could be the same for frequent and infrequent patterns, meaning frequent sequences are not necessarily more stored than infrequent ones.

And then in this case, in our opinion, what is planning, what is programming, and what is executing?

Ben Parrell

I think that we haven't given good definitions for planning and programming.

I intended 'planning' to refer to things that can be learned or optimized over. And then, programming would involve executing those things in context. You could start to optimize over larger or smaller units, and I think each of those can form a motor plan as they are predictable sequences—they are produced over and over again.

I think this connects to the point you raised about coarticulation, Marina: there is more coarticulation and more consistently within a word than across words. I think there are two different types of coarticulation: one involves optimizing a motor plan, and the other involves optimizing that plan in context.

I think it is very similar to your model, Pascal, right? The controller has a cost function that applies to a larger sequence, but the individual items—where Bayesian estimation is used in the current version of GEPPETO, right?—are also optimized over time and learn to be produced well. I don’t see them as conflicting, but to me, there is a distinction between planning a constituent and programming it within its context. However, I’m not sure how this relates to reaction time.

[3] Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. Journal of cognitive neuroscience, 22(7), 1504-1529.

Frank Guenther

Our interpretation is similar to yours. When we train participants to say these words, we notice a couple of things that change significantly over the course of training. Initially, it seems as if they produce one phoneme at a time, so producing the first phoneme doesn't take long. But, by the end, they are producing a larger chunk of several phonemes. At that point, I would expect production to take slightly longer because, initially, it is a sequencing problem—sequencing through all the phonemes. The first phoneme can be produced quickly, but sequencing through the rest takes more time. In my view, the second process involves reading out the entire chunk, but if the chunk is long, it takes more time to initiate.

Carrie Niziolek

Perhaps I can expand on this by considering the extent to which reaction time reflects some level of pre-activation of the unit. This is influenced not only by whether the unit is stored or constructed online but also by the extent to which it has been activated by prior context. For example, if a word has been repeated many times during the experiment, its activation state will likely differ from that of a newly introduced word. To the point where, at least temporarily, the distinction between something already in your lexicon and something that could become part of your lexicon during the experiment might blur. I think we've struggled with the question of how accurately reaction time reflects underlying processes. The general idea is that reaction time can be a powerful measure, but it's difficult to determine the level at which the timing or delay is coming from.

Ben Parrell

The ambiguity in reaction time is, I think, one reason why we need multiple approaches to study planning. Each approach has its own limitations. The neural data, different psychophysical data, and reaction time data all start to point to the same thing, which I think is good.

Marina Laganaro

That is indeed why we look at reaction time and brain signals.

Ben Parrell

I think this is important both for advancing our understanding and for practical applications. And while a healthy system may plan everything seamlessly, these levels can break down in different ways in neurological disorders. I think this is an interesting area to explore, but it requires clear definitions. For example, what distinguishes programming from planning? I'd be happy to call them 'Planning One' and 'Planning Two'—as long as, like you said, we acknowledge that we use different terms. If we can place them within a common framework, we might move past these differences. But clearly defining what we mean is crucial; otherwise, we risk using the same words in different ways across different contexts.

Cécile Fougeron

One follow-up question could be: is controlling the same as planning? Perhaps we could define 'Planning One' as planning, and 'Planning Two' as both planning and programming. Then, we might introduce 'Planning and Controlling' as another way to distinguish between the two.

Frank Guenther

Motor planning basically involves going from some abstracted non-motoric thing into a time series of variables that are going to be controlled in a way. And to me, that stage closely resembles motor control in industrial systems, where variables like temperature are regulated, for example. Control theory has many examples of this, where a system has a clear-cut target that you are trying to reach, that varies over time, and variables that you can adjust to reach the target. Planning, on the other hand, is something different than that. It involves going from something abstracted, more language oriented, into a motor trajectory—essentially a sensorimotor trajectory. To me, that is the key difference.

Ben Parrell

Or, since not all of us work with trajectories—the current version of GEPPETO, for instance, doesn’t incorporate them, nor does task dynamics—it could instead involve developing a feedback controller, which is then used by the controller in the plant to execute movement, though the underlying idea remains the same.

Frank Guenther

What I mean by trajectory is a time series of targeted things, and I think the task dynamic model does have trajectories in this way.

Ben Parrell

But the time series is emergent, right? Based on the controller as a function of time.

Frank Guenther

But it's something that varies continuously.

Pascal Perrier

Can we say that what Ben initially referred to as planning is different from specifying a sequence of goals? These goals could take the form of trajectories or not, but could we describe them as a kind of sequence of ideal goals? Interestingly, you distinguished between 'Planning One' and 'Planning Two'—where 'Programming,' in your view, is actually planning in context.

Ben Parrell

Yes.

Pascal Perrier

So, planning in context means taking into account the conditions under which execution will occur, including some kind of physical reality. So, is planning more about specifying a series of goals—whether as a sequence in time or discrete goals?

Ben Parrell

Yes, as long as those goals can be optimized in some way to achieve the desired outcome.

Pascal Perrier

At this stage, would you agree that the way linguistic units are reflected in this part of the planning could be a kind of duration or sequence over which optimization occurs? This could possibly define levels, because we could imagine optimization occurring at the level of a syllable—perhaps the strongest level—followed by optimizing a sequence of syllables. Starting from this optimized pattern, further optimizations could incorporate prosodic constraints and ultimately longer utterances. Would you agree with this suggestion?

Ben Parrell

Yes.

Pascal Perrier

Carrie, not?

Carrie Niziolek

No, I do. Actually, I'm glad you brought up goals because I think that provides another piece of evidence we can use. The way I think about the planning level is that when we consider a goal that is transformed into a sensory space, that we can predict and compare to incoming feedback, the evidence suggests that this process occurs at the planning level rather than at the programming or execution level. In terms of the brain's response to hearing expected feedback, it seems to be that the expectation is at the level of planning more so than the level of the execution—in terms of the unit being predicted.

Most of my evidence comes from the neural regions that seem to underlie the prediction, which concern more, for example, the ventral premotor cortex rather than motor cortex, or even higher-level planning areas like the IFG. Those seem to be, at least for the auditory cortex, the source of sensory prediction. We could discuss this further in the impaired speech roundtable, but if that area is lesioned or if we examine a brain where both regions and their connectivity can be analyzed, it appears to be linked to the premotor region.

Ben Parrell

I would add that this might be unique to the auditory system. I think that the auditory system is inherently more informative about our high-level goals. Whether those are auditory targets or constriction targets, for example. And in contrast, the somatosensory system is likely more informative about low-level execution.

Carrie Niziolek

I agree with that.

Frank Guenther

Well, auditory feedback comes in at several levels. For example, people vary over the course of the day in whether their first formant is high or low. If their formants are perturbed at a high point versus a low point in the day, they adjust based on what they were doing at that specific time. They don't have an idealized target. To me, this suggests a projection from motor areas to sensory areas that conveys what should be happening based on the motor commands. Then, at a higher level, there is monitoring of phonemic errors, for example. So, there are very different kinds of use of auditory feedback at different levels and I think that some of it is quite low level.

Ben Parrell

Even at that low level, auditory feedback still relates to an abstract target even if it is relative to the current action, because it's not about how you can actually move your body, right? Somatosensory feedback is intrinsically tied to how the movement is executed, whereas auditory feedback pertains to a high-level consequence.

Frank Guenther

Right, and it's also much faster.

Pascal Perrier

I am convinced that speech is primarily auditory. When it comes to speaking, I believe the first thing we do is attempt to reach an auditory goal, because we communicate with acoustics.

At the same time, since we speak a lot, we associate both modalities. So, under normal conditions, I don’t see why audition should necessarily be considered higher-level than somatosensation. Actually, in experiments with Jean-François Patri, we demonstrated that people are able to categorize vowels based on somatosensation. This suggests that, under normal conditions, the link to phonology appears to be similar for audition and somatosensation.

Ben Parrell

What I mean to say is that the auditory signal is informative about the auditory state, but it is not informative about how we actually need to move our articulators to achieve that state. By contrast, the somatosensory system is directly tied to the current state of the articulators.

Frank Guenther

I think that is true if we consider muscle spindles, but tactile information is quite remote from the motor space.

Pascal Perrier

Why don't we see that differently?

When we move or speak, we have an impact on our world, which in turn affects our perception—auditory, somatosensory, and tactile.

Somatosensory perception can involve muscle spindles or mechanoreceptors—those are all types of sensations. So why should we prioritize one sensation over another? Audition, after all, is just another consequence of our movements—just add somatosensation in. What makes a difference is the fact that we communicate through acoustics. We know that the person we are speaking to extracts information from it, but otherwise, those are all sensations.

Carrie Niziolek

I think that, depending on the state of the system, the same motor command—even with largely identical somatosensory feedback—can produce different acoustic signals. Factors such as dehydration late in the day or having a cold may introduce differences. I do think that over the course of the day, various factors may cause differences in execution or, more specifically, in acoustic realization—even if the underlying planning remains the same.

Audience

Is the prediction and the prediction error always online?  So, is there a constant and continuous speech predictive monitoring?

Carrie Niziolek

I think so.

Frank Guenther

At some level, yes. You can modify a formant anywhere in an utterance, and people will correct for it. The delay seems to be consistent—always around 100 milliseconds or so. So, I believe this process is online. However, auditory feedback is not used at just one level. Beyond immediate corrections, there is likely additional processing after a sentence is completed, which may lead to adjustments at either the planning or motor level. Auditory feedback operates at multiple levels throughout the speech process, serving different functions at different stages.

Ben Parrell

Yes, great.

Cécile Fougeron

Carrie, you were saying that execution may vary throughout the day because it accounts for physical constraints. But are these physiological and mechanical constraints modeled at the execution level or within programming?

I don't think these constraints are handled only at the execution level; I would also see them at the programming level. In this sense, programming would involve control while accounting for the physical world, with information then sent to the muscles for execution.

Carrie Niziolek

I agree with that. I think that is the difference between reacting to an unexpected perturbation on the fly versus adapting to longer-term changes. For example, if you detect that you're tired or dehydrated, you can learn and adjust within a minute, potentially modifying things at the programming level for a longer-term change.

Audience

There have been many comments about motor programs and how we reuse frequent sequences—words, syllables, etc. However, we all know that when words are placed in an utterance, they are produced differently depending on their context. In this sense, we almost never, if ever, reuse the exact same spatiotemporal or spectrotemporal sequence. This raises a fundamental question about the definition of a motor program: What is it that we actually reuse or store when we speak? What do you think a motor program actually is?

Ben Parrell

Do you want my engineering answer? I think an optimized controller.

Audience

Can you elaborate on that? For instance, you could have a target at a single point in time that gets optimized at one level.

Ben Parrell

Well, in a feedback control system—like the GEPPETO model Pascal showed earlier—you can think of the controller as a time-varying cost function. If you have learned to minimize that cost function, you may also have constraints that are contextual and change that cost function in a particular context. I think what I would call programming is this additional cost function.

Audience

So, you are calling motor programming the cost function?

Ben Parrell

It’s the end result of the programming process.

Frank Guenther

I think of the motor program more as a dynamical system that you activate. Even if the motor program itself may not change, its actual instantiation will vary depending on factors like the position of your tongue at onset, or whether you're about to stress the next word. So, there are certainly aspects that can be adjusted—some of which might even be consciously adjusted. But most of it, if you're specifying targets in a higher-level space and you're producing them in a lower-level space, there is redundancy. The DIVA model, for example, won’t produce the same outcome every time; it depends on the starting conditions, even if the underlying motor program itself remains unchanged. The resulting movement may look fairly stereotyped, but the details are actually different for each production. That’s because we are not simply storing just a set of muscle activations over time, but rather something more flexible—more like a dynamical system. I think this idea aligns well with both the DIVA model and the Task Dynamic framework, if I’m not mistaken.

Ben Parrell

Yes.

Pascal Perrier

I understand your question very well, because I ask myself the same thing. I have to admit that I don't fully understand what exactly a motor program is—and I'm not sure the term is even clearly defined. At the same time, I have the impression that the process I described—what you refer to as optimal feedback control—does not systematically occur in this long process, starting from nothing. At this stage, I would say that a motor program is perhaps best understood as a kind of initial configuration—for example, to produce a syllable—that serves as an initialization for the optimization process. This makes the optimization process short and not very demanding. Otherwise, if we required a fully specified motor program for every syllable in every context—clear speech, unclear speech, fast, slow, loud, whispered—that would be highly inefficient and clearly not economical.

Ben Parrell

And I guess I would add that I think, to bring all of those ideas together, if you think about a motor program as seeding a dynamical system, like in a neural state space, and that neural state is gonna be influenced by sensory feedback, and it's gonna be influenced by their context, and that system is gonna evolve, right? But seeding that initial state of a dynamical system could be part of the motor planning process.

## Units

Louis Goldstein

Somewhat surprisingly to me, I actually don't disagree with anything that's been said so far. I don't have much to offer with respect to distinctions between planning and execution, other than the issues that have been brought up, which were interesting and useful.

Just to play devil's advocate, I'd like to throw something else into the mix. There was a question about which units are encoded—and I think if we've learned anything from large language models, it's that the answer is: all of them. If you can think of a unit—any unit—that has even the faintest argument in its favor, the models know about it.

And you can find evidence for it in the actual representations that LLMs give you for a given acoustic sequence. Those representations aren't necessarily what's going on in the brain—but there is evidence that there is a relationship between the representations the models find at various layers and neural representations in different areas of the brain[4].

So, you know, the discussions here have been good—but I think the issue of *which* units might not be the most productive one to focus on. Honestly, I think the answer is: you can find a representation for any of them, and they're probably all useful under different conditions.

Frank Guenther

Louis, just to follow up on that—the way I see it, not every unit is in every part of the brain and there are parts of the brain that don't seem to care about phonology that are definitely involved in motor control. And then there are other parts where it seems syllables are more frequently represented than other things. So, if we consider the overall system, then I completely agree with you. But if we consider a given brain area, I think there is some hope in finding units—even if they're not single, simple ones.

Louis Goldstein

Yes, I agree with that.

Marina Laganaro

I think we all agree that the system is flexible—you can plan with syllables, smaller units, or larger ones. But what makes us resize? Why do we sometimes plan in larger units and other times in smaller ones? What's behind that? What's constraining this?

Frank Guenther

[4] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunber, C. Pallier, J-R King (2023). Toward a realistic model of speech processing in the brain with self-supervised learning. arXiv:2206.01685v2

The language system is feeding information to our motor system, and that is of course one of the factors. I think the question becomes more interesting when we think about a person with a disorder, or situations where they have to slow their thinking because the production system can only go so fast and they can only hold so much in mind. The language system itself is capable of so many things, and each time you use it, it may give you a different chunk. Sometimes it is a small chunk, and other times it’s more thought out—like a full phrase. But that varies a lot across individuals, and it varies even more in cases of disorder.

Ben Parrell

I’d also add that the units we observe really depend on how we probe them. If we are really planning over all these levels, it'll look like we're planning over different levels depending on how you ask your question. I think that’s what the adaptation work Frank and I were discussing earlier points to: sensorimotor adaptation tends to generalize broadly unless you create a condition where it won’t. So, if you only look at that narrow condition, you’ll draw one conclusion—but if you probe more broadly, you might see something quite different. That’s why it’s so important to consider how you’re asking the question when you interpret the results.

Louis Goldstein

The question of which unit is being used is going to be sensitive to the conversational context, right? It depends on turn taking, and the temporal aspects of turn taking, as well as actual meaningful aspects of turn taking.

Pascal Perrier

Yes, Louis, I wanted to say that too—that we communicate with people, after all. And regarding the issue of the syllable: the physical system is actually one well-known reason for focusing on syllables. The opening of the mandible generates variation in the amplitude of the signal that helps parsing. And in the brain, the theta oscillation enables us to synchronize with the syllabic rhythm. So for these reasons, the syllable should definitely be considered a level of programming.

# Round table 2: Neurotypical versus impaired speech

*Doris Mücke (moderator), Louis Goldstein, Frank Guenther, Antje Mefferd, Caroline Niziolek, Pascal Perrier, Wolfram Ziegler*

## Summary

*The discussion in round table 2 addressed issues of modelling atypical versus typical speech, the continuity between the two, compensatory mechanisms and optimization and finally the inter-speaker variability.*

***Modelling typical versus atypical speech*** *When modeling atypical speech, there was broad agreement that models should aim not only to capture static end states, but also to account for the developmental trajectories that lead to them. A dynamical systems perspective can be helpful in this regard. Adaptive models—those that learn over time—offer a promising way to capture individual differences and diverse outcomes.*

*Given that errors are present even in typical speech, it was discussed whether there is a continuum between typical and disordered speech. Some argued that disordered patterns often reflect exaggerated versions of typical variability, meaning that there is a way in which the analysis of disordered systems can reveal what the typical system does when it is under duress. Others cautioned against assuming a linear continuum, pointing instead to distinct developmental trajectories or "bifurcations" that may lead patients down different compensatory paths. Still, many agreed that meaningful analogies can be drawn between how both typical and disordered speakers flexibly use their systems.*

*Related to the discussion that errors are observed even in typical speech, there was partial disagreement about whether language/speech is non-optimal. Some argued that speech errors do not necessarily reflect a lack of optimization. Instead, optimization might be better understood in terms of resource constraints and trade-offs. Given the brain's limited capacity, the system may still be optimal with respect to goals like energy efficiency or information transfer—even if it sometimes fails. "Optimal" in this context does not mean perfect, but rather the minimization of cost with respect to communication demands.*

***Compensation and optimization*** *The language system adapts over time through compensatory mechanisms, optimization under constraints, and probabilistic learning based on prior experience. When one region is impaired, speakers do not simply fail—they often adjust by shifting articulatory strategies. These compensatory paths are not uniform; they can diverge unpredictably and differ across individuals. Optimization can be seen as an (emergent) outcome of adaptive systems, whether the system is working around a constraint or learning speech for the first time during development. It was agreed that there are multiple paths to optimization, and this depends on anatomical and cognitive characteristics of the speaker. Optimization may include changing planning strategies, such as the size of planning units. It was also said that while optimization is a useful engineering metaphor, real behavior might be better captured by probabilistic, experience-driven models.*

***Inter-speaker variability*** *Speech production strategies are shaped by anatomical, perceptual, and experiential differences, leading each speaker to arrive at a slightly different solution for what is optimal for their own system—even if that solution is not necessarily optimal across a broader population. For instance, speakers with dysarthria may rely more on jaw than tongue movement to achieve a similar overall constriction degree. Individual differences also emerge in the size of the buffer—that is, how much linguistic material is passed along to the motor system—and in the perception of effort, with some speakers perceiving a task as minimally effortful while others experience it as more demanding. Developmental factors and accidental learning play a role as well: certain articulatory patterns may be established early in life, such as during a period when the tongue is proportionally large compared to the head. These early-acquired patterns may persist into adulthood. Notably, variability does not generalize uniformly across the sound system—a speaker may show high variability for one segment while remaining stable on others. There was an agreement that the development of speech systems across lifespan is important to understand and to model typical and atypical speech system behavior.*

Doris Mücke (moderator/discussant)

What problems arise from the fact that most language production models have been developed on the basis of healthy speech, rather than pathological speech.

A common goal of kinematic studies on disordered speech is to identify speech motor impairments—essentially to capture, or quantify the negative impact on speech function. This includes precision, speaking rate, or speech intelligibility. However, systematic quantitative assessment of these impairments—particularly movement disturbances—remains a major challenge. For instance, if you have tried to fit kinematic contours from electromagnetic articulography (EMA) into a Task Dynamics model, you will know exactly what I mean.

The speech movement patterns of speakers with speech movement disorders often deviate substantially from those of healthy speakers. For instance, we discussed the role of stiffness for the speech system—but in cases of impaired speech, we frequently observe multiple velocity peaks, which makes it very difficult to determine whether stiffness as a reliable control parameter is even present.

Speech pathologists, phoneticians, and phonologists currently cannot fully benefit from each other's work although their overall research goals do overlap considerably. One of the main challenges is that many speech production models have been developed with a focus on healthy speech. This makes it difficult to connect insights across disciplines. Perhaps the DIVA model is an exception to some extent—but that's something we can discuss further here.

In the figure 2, I am showing kinematic contours taken from a standardized DDK task. We could debate whether DDK is truly “speech,” but on top you can see a simple /papapa/ sequence from a healthy speaker. Now, compare the healthy speaker with data from a patient with essential tremor while the deep brain stimulation (DBS) is deactivated. You can see a strong increase in variability—this is what I mean by multiple velocity peaks. When DBS is activated for the same patient, you can see highly irregular syllable cycles, which sometimes break down entirely. These patterns are hard to account for in the existing speech production models.

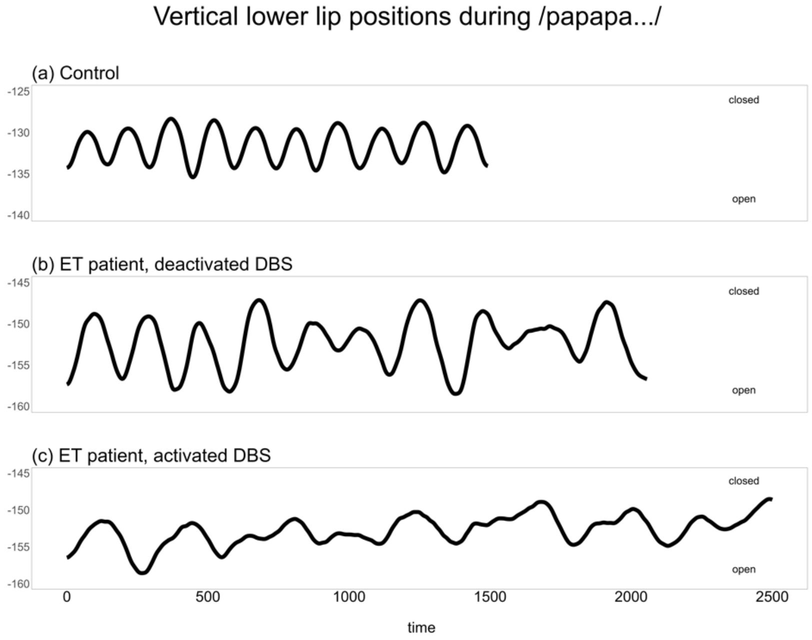

*Figure 2: Vertical lower lip movements during the DDK task of /papapa/ produced by a healthy speaker (a) and patient with essential tremor with deactivated (b) and activated (c) deep brain stimulation (Mücke et al. 2024)*[5]*.*

Frank Guenther

I think there are a couple of different but related questions here. One question concerns how best to understand the disorder. And for me, that begins with understanding the healthy system—what it is doing normally, and what is going wrong when things break down. That said, there are plenty of cases where you're studying a disorder and a different measure is important that people don't use for healthy speakers.

I think there is always going to be a combination of disorder-specific factors. But when it comes to understanding a disorder—if that's the goal—I think a logical approach is to first understand the basic, healthy system, and to then figure out how it goes wrong.

With DIVA, the idea is to make it easy to selectively impair different components of the system. When we do that, we can observe how brain activity and movements change, which helps us better understand what's happening in disordered speech—and, conversely, it also helps improve the model itself.

For this reason, we often study disorders because they provide insight into the normal system. You learn about how something works by seeing how it breaks down. So I wouldn't advocate abandoning the healthy model as a basis for understanding the disorder. But I also personally think that we should not only study healthy speech in order to better understand that first.

[5] Mücke, D., Roessig, S., Mefferd, A., Thies, T. & Hermes, A. (2024). Challenges with the kinematic analysis of neurotypical and impaired speech: measures and models. *Journal of Phonetics, 102*, 101292. https://doi.org/10.1016/j.wocn.2023.1012

Wolfram Ziegler

I'd like to respond to that point. You mentioned that these models are based on typical speech—but actually, many of them have historically been grounded in neurological data. For instance, ideas about auditory–motor interaction go back as far as Wernicke. So, in a sense, these models are already a blend—a mixture from pathological cases and more modern techniques.

Take fMRI, for instance—it only emerged in the 1980s or 1990s, and it has had a significant influence on current speech models. But many of these models can actually be traced back to much earlier clinical frameworks. There was a longstanding discussion in cognitive neuropsychology about how to infer processing modules from patient data—and mostly from rare or unexpected cases. That's where components like lexical reading or non-lexical reading, for example, originated.

The issue was that this was based only on patient data and there was a proliferation of such modules; with each new patient, you had a new module. So, one of the key tasks of contemporary modeling is to correct that—essentially to ask how much of it actually reflects the mechanisms underlying normal speaking.

So, the discussion then was: when you observe a brain lesion in a patient, do you assume that the lesion removes a specific processing module and leaves everything else intact? Or do you take the position that the lesion disrupts the functioning of the whole system more broadly? These two positions were very much confronting each other. And I think the models we are working with today are built from a mixture of pathological models and models using modern techniques.

Louis Goldstein

I think we need to know a lot more about the phenotype of each kind of disorder.

While we often frame disordered speech as something going wrong in a system developed for typical speech, it's also true that disordered behaviors can be fairly prototypical of a given disorder—they are in their own stable states. Ideally—and I say this with caution, since I know very little about most disorders—we want a dynamical system that is abstract enough to account for both. That is, a system with attractor states we identify as "typical", but then, under certain conditions there are other semi stable states that the system shows that arise from other kinds of inputs to the system—or other kinds of control parameters, if you will.

Some of those parameters are neural—that would be the function of the brain—others might be speech rate. This is what I would view as the optimal relation. To do that, we need a much more detailed understanding of the specific dynamics of speech in different types of disorders. And then the question becomes: what kind of nonlinear dynamical system allows us to get from one to the other? That might be aspirational at this point—but that's what I would imagine is ultimately the most useful approach.

Frank Guenther

I largely agree with that. And I'd like to add that we need to keep something else in mind: if, for example, if someone has cerebellar damage and develops ataxia, we cannot simply 'damage the cerebellum' in a model and expect it to replicate the patient's behavior. That's because people compensate. So if we want the model to reflect what actually happens in a patient, it needs to include compensatory mechanisms. And that makes things much more difficult—but certainly not impossible.

Carrie Niziolek

I think that’s a really helpful way to think about it. I do think there are some disorders that can be characterized by damaging a specific region—maybe with the addition of compensatory mechanisms. But in other cases, what we see in disordered speech are simply more extreme versions of patterns that already exist in what we call typical or healthy speech.

Take speech errors, for example. In aphasia—which I know more about—we used to think the problem was primarily a misselection of units at a high level, like swapping one syllable for another. But when you look more closely at the kinds of errors people with aphasia make, you see blended errors that actually mirror what we observe in typical speech errors.

So, there is a way in which the analysis of disordered systems can reveal what the typical system does when it’s pushed to an extreme or under duress. I think that kind of insight can be just as revealing as pinpointing a region that is damaged or has parameters set differently.

Antje Mefferd

I can actually add to that a bit. When we look at speakers with dysarthria, I’m often more interested in examining each articulator individually, because articulators can be differentially affected. For example, the tongue is often more impaired than the jaw (e.g., [6]).

If we only look at constriction degree, we risk missing important differences between speakers with dysarthria and healthy controls. A similar constriction degree might be achieved in both speaker groups, but in speakers with dysarthria it could be primarily driven by jaw movement, whereas healthy speakers may rely more on tongue movement (e.g., [7]).

[6] DePaul, R., Abbs, J.H., Caligiuri, M., Gracco, V.L., Brooks, B.R. (1988). Hypoglossal, trigeminal, and facial motoneuron involvement in amyotrophic lateral sclerosis. Neurology, 38, 281-283.
Langmore, S., & Lehman, M.E. (1994). Physiological deficits in the orofacial system underlying dysarthria in amyotrophic lateral sclerosis. Journal of Speech and Hearing Research, 37, 28-37).
Mefferd, A.S., Lai, A., Bagnato, F. (2019). A first investigation of tongue, lip, and jaw movements in persons with dysarthria due to multiple sclerosis. Multiple Sclerosis and Related Disorders, 27, 188-194.
Mefferd, A.S. & Dietrich, M.S. (2019). Tongue- and jaw-specific articulatory underpinnings of reduced and enhanced acoustic vowel contrast in talkers with Parkinson's disease. Journal of Speech, Language, and Hearing Research, 62, 2118-2132.
Mefferd, A.S. & Dietrich, M.S. (2019). Tongue- and jaw-specific articulatory underpinnings of reduced and enhanced acoustic vowel contrast in talkers with Parkinson's disease. Journal of Speech, Language, and Hearing Research, 62, 2118-2132.
Yunusova, Y., Weismer, G., Westbury, J.R., Lindstrom, M.J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. Journal of Speech, Language, and Hearing Research, 51, 596-611.

[7] Mefferd, A.S., Lai, A., Bagnato, F. (2019). A first investigation of tongue, lip, and jaw movements in persons with dysarthria due to multiple sclerosis. Multiple Sclerosis and Related Disorders, 27, 188-194.
Rong, P., (2019). The effects of tongue-jaw coupling on phonetic distinctiveness of vowels in amyotrophic lateral sclerosis. Journal of Speech, Language, and Hearing Research, 62(9), 3248-3264.

But if the system is constainedconstrainedconstained —for instance, by fixating one articulator with a bite block—you might see a similar compensatory behavior in a healthy speaker and someone with dysarthria. It's about finding the best solution given the current constraints of the system[8].

Pascal Perrier

You must have had a specific idea in mind when you set up the DBS. What brain region did you choose to stimulate, and why did you choose to stimulate this region of the brain?

Doris Mücke

It was a stimulation of the ventral intermediate nucleus of the thalamus (VIM)—a deep thalamic structure. The idea is that in patients with essential tremor, speech often deteriorates significantly after surgery. While gross motor control is typically improved under DBS, the speech sounds very slurred for many patients after DBS implantation. This type of stimulation-induced dysarthria is specific to patients with essential tremor stimulated in the VIM region; it's not something we usually see in patients with Parkinson`s disease stimulated in the STN region, for example.

In stimulating the VIM, the primary goal is to suppress the tremor, because patients often reach a point where they can't even hold or grasp objects properly. So, the main purpose of stimulation in this case is to improve gross motor control and stop the tremor signal, since essential tremor causes widespread shivering in whole parts of the body. The deterioration of speech is mainly a side effect of the stimulation.

Frank Guenther

Well, they were doing that long before we understood the effects of stimulating that area. With VIM, the main reason is that the tremor is linked to a cerebellar circuit—and VIM is the part of the thalamus that receives input from the cerebellum. So, by stimulating VIM, you're essentially interrupting the loop that gives rise to the tremor. Now we understand why it works, but I don't think that's the original reason they started implanting electrodes in VIM, though, I'm not entirely sure.

Doris Mücke

Stimulation might also affect the motor fibers of the internal capsule located laterally to the VIM, so that might be part of the issue.

Pascal Perrier

The reason we can't explain this result with current models of speech isn't necessarily because they're based on healthy subjects, but rather because they are, at their core, functional models. There's no physics in them. We model the function of the overall system, but we don't truly model how information physically goes into the brain, between different brain regions such as from the cerebellum to other areas.

Frank Guenther

[8] Mefferd, A. & Bissmeyer, M. (2016). Bite block effects on vowel acoustics in talkers with amyotrophic lateral sclerosis and Parkinson's disease. Journal of the Acoustical Society of America, 140, 3442.

In our model[9], we specify every brain region—each component is associated with a specific location in the brain (see Figure 4).

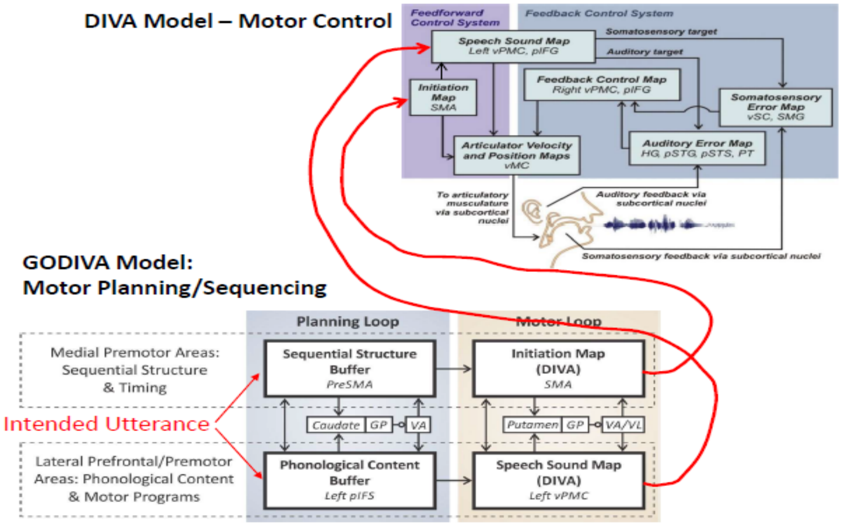
DIVA Model – Motor Control
Feedforward Control System
Feedback Control System
Speech Sound Map
Left vPMC, pIFG
Somatosensory target
Auditory target
Initiation Map
SMA
Feedback Control Map
Right vPMC, pIFG
Somatosensory Error Map
vSC, SMG
Articulator Velocity and Position Maps
vMC
Auditory Error Map
HG, pSTG, pSTS, PT
To articulatory musculature via subcortical nuclei
Auditory feedback via subcortical nuclei
Somatosensory feedback via subcortical nuclei
GODIVA Model:
Motor Planning/Sequencing
Planning Loop
Motor Loop
Medial Premotor Areas: Sequential Structure & Timing
Sequential Structure Buffer
PreSMA
Initiation Map
(DIVA)
SMA
Intended Utterance
Caudate
GP
VA
Putamen
GP
VA/VL
Lateral Prefrontal/Premotor Areas: Phonological Content & Motor Programs
Phonological Content Buffer
Left pIFS
Speech Sound Map
(DIVA)
Left vPMC

*Figure 4. Sketch of the GODIVA and DIVA models (slide from the morning talk, see references in Footnote 9)*

Some of those mappings are probably wrong, but having that anatomical grounding gives us something concrete to test and refine. We improve the model over time by generating predictions, testing, etc.

Pascal Perrier

Yes, sure—though my point was a bit different. What I meant to ask is: does your model simulate how information flows from one brain region to another? For instance, can it account for the presence of noise in the signal? For example, can you take into account the fact that there is noise in the signal, and that this noise can be different according to the patient?

Frank Guenther

To some extent, yes. The model includes noise sources in its pathways—for example, we can selectively impair the cerebellar pathway, and we should observe something that resembles what we see with VIM stimulation. We haven't yet run those experiments or simulations, but that's the direction we're aiming for.

What I'm advocating for is more discussion around not just functional components, but also where these components are actually located in the brain. Because ultimately, I think our best understanding will come from what the brain is actually doing—what we have measured and understood. The brain is where we

[9] Guenther, F.H. (2016). Neural Control of Speech. Cambridge, MA: MIT Press.
Tourville, J.A. and Guenther, F.H. (2011). The DIVA model: A neural theory of speech acquisition and production. Language and Cognitive Processes, 25, pp. 952-981. PMCID:PMC3650855
Bohland, J.W., Bullock, D. and Guenther, F.H. (2010). Neural Representations and Mechanisms for the Performance of Simple Speech Sequences. Journal of Cognitive Neuroscience, 22 (7), pp. 1504-1529. PMCID:PMC2937837

should be looking—and while we start with informed guesses to guide our search, those guesses improve over time as we measure and understand more.

Especially in speech, I think we already have enough information to build models that are anatomically specific—where we can say which brain regions are responsible for what. Language may be less amenable in this respect, but at this point, I don't think there is any reason not to start associating components of our models with specific brain areas.

Marina Laganaro

I agree with Carrie's comment that there may be a continuum between typical and atypical neuromotor speech disorders. You mentioned aphasia as an example, and I completely agree—it has been demonstrated in language models that aphasia can be simulated by introducing noise or increased competition, which supports the idea of a continuum.

This brings me to another issue: it has been acknowledged that the language system is non-optimal. We all produce errors, and it's not optimal in the sense that we often select the first available word just to communicate, even if it's not the best one.

But what about speech? Do you also think the speech system is non-optimal? And how can we model this non-optimality in speech production models? After all, we do make speech errors quite frequently.

So in light of this idea of a continuum, what happens in the case of motor speech disorders?

Frank Guenther

In my view, just because the system makes occasional mistakes doesn't necessarily mean it's not optimal. Given the limited amount of brain space, these errors might not be the best the system can do but it may be optimal in terms of information transfer rate, for example. The system may still be optimal in terms of energy expended. Those are the kinds of things that our brains automatically try to minimize, or maximize in the case of the information transfer.

Marina Laganaro

At least it's not perfect—would you agree?

Frank Guenther

Yes—but that's because we've only got a limited box on our shoulders. That's part of it, at least.

Cécile Fougeron

I'd like to expand on this idea of optimization. I think the reason we want optimization in our models is because we approach the problem from an engineering perspective. But then how do we account for speaker variation? Are we saying that some speakers are simply better optimizers than others?

Take someone with a motor speech disorder, for example—despite all the system-based constraints, they will still find a way to speak, perhaps by resizing the planning unit. More generally, any speaker might reduce their flexibility and shift to syllable-by-syllable planning when faced with challenging speech conditions—and that, too, is a form of optimization.

So maybe, rather than building models that define a single path for optimization, we should aim to reflect the flexibility of the system. There may be multiple ways to reach a communicative goal under different constraints. Patients make use of that flexibility, and we need models that can somehow deal with these different paths.

Antje Mefferd

I think a lot about inter-speaker variability because I deal with it in my data quite a bit. To some extent, we all have slightly different speech motor systems to work with—differences in anatomy, in perceptual acuity, or even in how we perceive effort. I think these differences contribute to the inter-speaker variability we observe and it makes it difficult to clearly define a disorder. Sometimes, control speakers will show patterns that overlap with those exhibited by speakers with mild dysarthria. I study individuals who are still intelligible, and in milder cases, there can be a lot of overlap in performance. Everyone arrives at a slightly different solution depending on what is optimal for their motor speech system. That's why we need large-scale studies to identify subgroups within our "normal" control group, so we can better interpret our findings of speakers with dysarthria.

Pascal Perrier

I fully agree. First and foremost, "optimal" doesn't mean perfect—it is just a minimization of the cost. And that cost, as you rightly pointed out, is speaker-dependent. I also want to mention something, Cécile: I don't actually believe in optimization of speech. I'm just using that term because, from an engineering perspective, it is currently the most workable approach. But personally, I prefer another proposal—one where we think in terms of probability. That is, we aim for the most likely motor command pattern or strategy to help us reach our intended goal.

So we rely more on an experience in which we have made different trials. And based on those trials, we've learned that if we want to achieve a certain goal, the most probable motor command pattern is this or that one. Personally, I believe more in that kind of probabilistic learning than in optimality. But optimality is a useful engineering approach for modeling it.

Cécile Fougeron

Does this means that if we want to model variation and flexibility in the system, we have two main options: either the goals themselves can change—maybe we don't always have the same goals—or we don't have the same ways to get to the same goal?

Pascal Perrier

There are different experiences. What I'm proposing is, first of all, that optimization is speaker-dependent. But more than that, if we take the view that motor behavior is shaped by experience—used to generate probabilistic patterns—then it follows that each speaker's experience is different. Life is different, our vocal tracts are different, our brains are different—so the experiences are different. As a result, the probability distributions we build up for selecting motor patterns to reach a given phonetic or linguistic goal is simply different from person to person.

Wolfram Ziegler

When a patient has an acute stroke, they learn to cope with the resulting impairments. For example, in cases of velar insufficiency—where most of the air escapes through the nose—we observe different compensatory strategies. Some patients open their mouths wide to balance nasal and oral resonance, while others close their larynx to slow down the airflow. These compensations vary: some are helpful and would be supported by a therapist, others are miscompensations in a way.

The point is, we don’t yet fully understand how these different compensatory paths emerge. That’s why I wouldn’t entirely agree with the idea that there is a simple continuum between typical speech and disordered speech. Instead, there are different bifurcations during the history of a disorder—some patients go down one route, others another—and we don't exactly know how that works.

Carrie Niziolek

I don’t think everything follows a continuum—I agree with that—but I do think there are analogies we can draw. Take, for example, the articulation of a rhotic sound: some people do it one way, other people another, and some switch between strategies depending on the context. There are analogs here in how individuals make use of the systems they have, whether you describe it as "optimal" or simply what’s easiest or most habitual.

A couple of other thoughts came to mind as we were talking, though the conversation may have moved past them. First, I think language is another differentiating factor among people that may influence strategies at the motor level. For instance, someone might have a different-sized buffer—how much they pass along to the motor system—which I think is another input.

Another thought relates to the brain regions that underlie different processes and compensatory mechanisms, in the context of DBS. Some individuals have to try to compensate effectively for a new neural signal or the perturbation to what's going on in the brain. I am more familiar with Parkinson’s than essential tremor, but even in Parkinson’s, some patients improve while others worsen after DBS. I think we need a better understanding of whether the stimulation is having a normalizing effect on the circuit—as it’s intended to—or whether it’s actually removing or disrupting important information in that circuit. I think some of the models that try to explain the deficits observed with DBS need to start by asking: what exactly is the DBS doing in the first place? Without that understanding, it's hard to determine whether the effects we see should be interpreted as impairments. And again, the outcome may differ from person to person.

Louis Goldstein

We may have moved on while I was temporarily offline, but I wanted to add something about optimization and individual differences. Beyond differences in physical plants and brain structures, there are also differences in developmental trajectories. Some patterns get established early on—for instance, during a stage of development when the tongue is relatively large compared to the size of the head—and those patterns can be learned and kept. As a result, we end up with measurable differences across individuals, including in the amount of variability observed when producing certain segments.

And this variability doesn’t generalize across the different segments of the language. One person might produce a highly variable /s/ or /t/, while their other segments are not so variable. So that is presumably not only based on anatomy and brain structures, but also on some kind of accidental learning, which takes place over the years.

Frank Guenther

One thing to add is that by having adaptive models—models that learn—we have a better chance to capture these kinds of differences. For example, in the DIVA model, if you simulate a vocal tract where the jaw doesn't move well, the model learns to produce speech differently—it minimizes jaw movement. So, it is an adaptive model. Optimization is inherent to the structure of the model, but it is not explicit. I think the brain works similarly: the neural circuits naturally settle into minimum energy states. To capture that, I think adaptive models that learn over time are becoming more important for those distinctions.

Audience

I've spent many years working on stochastic models of speech motor control based on adult speech, and now I focus on infants from birth to age three who are developing autism. One thing that is missing—but was just mentioned—is the importance of studying infant development, where many of the constraints on production and control actually begin to emerge. Autism is present from birth, but many of the constraints on these emerging systems don’t become apparent until the second year of life.

Many aspects of speech production are impacting those infants because of the robustness of biological canalization of development, but many of them—particularly those that rely on interaction and contingency—are impaired. So I wonder whether we need to shift from thinking of speech production models as describing a single moment in adult function, mathematically or otherwise, to instead grounding them in how these adaptive skills emerge and evolve over time.

As Louis mentioned, it’s not just about capturing a fixed state—it’s about understanding how the scaffolding of dynamic systems develops across the lifespan. I’d love to hear thoughts on how we might incorporate infant development into our models, and how you can conceptualize that in mathematical and empirical frameworks.

Frank Guenther

I think those are great points. Right now, our model includes only very simplistic developmental stages, and that needs to be improved. It's easier to start modeling with adults—easier to understand, easier to run experiments with. That said, I completely agree: we absolutely need to be looking at what's happening over development and how those building blocks lead to eventual behaviors.

# Round table 3: Short-term adaptations and speech modes/styles

*Hélène Lœvenbruck (moderator), Frank Guenther, Antje Mefferd, Doris Mücke, Ben Parrell, Pascal Perrier*

## *Summary*

*In the third round table the mechanisms of speech modes/modulations are discussed along with inter-speaker variability and clinical perspectives.*

***Mechanisms of speech mode adaptation*** *Speech mode adaptation was described as both a voluntary, task-driven process requiring cognitive control and a habitual process shaped by repeated use. These adaptations are likely achieved through tuning at multiple levels of the speech production system. Global settings may interact with local motor plans and, perhaps predominantly, with motor programming. There was general agreement that speakers likely do not store distinct speech plans or motor maps for each speech mode. Rather, speakers become more proficient at certain speech behaviors simply through repeated use.*

*Clear speech, prosodic prominence, and loud speech are not governed by a single low-level parameter. Multiple dimensions (e.g., articulation, f0, duration, intensity) are simultaneously adjusted. The mechanisms may differ based on the modulation we are trying to achieve. For example, linguistic prosody seems to involve the planning level or earlier because it is part of the meaning—it is part of the linguistic message. In this case, it would likely involve the left inferior frontal areas and the premotor cortex. Multimodal parameters were also briefly mentioned. For example, in loud speech, visual cues like co-speech gestures become more prominent.*

*Whispering may be actually quite different from the other speech modalities that were discussed (e.g., loud speech, clear speech), because it uses a different laryngeal gesture. During whispering the laryngeal gesture for voicing is replaced with whisper, but the gestures for voiceless sounds remain unchanged. It may pose specific challenges in disorders like apraxia of speech.*

***Sources of inter-speaker variability*** *Constraints on speech modulation can be physiological, cognitive, or socio-communicative in nature, meaning that speaker variability can arise from both cognitive strategies and anatomical differences. In different situations, speakers may try to optimize different constraints on communication. Also, some speakers rely more on sensory feedback; others on predictions. Those that rely more on feedback may adjust their speech more. On the anatomical side, vocal tract differences also shape speech patterns and control. For example, many differences in control can be attributed to differences in palatal shape.*

***Clinical perspectives*** *The basal ganglia, which receives input from multiple cortical areas, is implicated in the adaptation of speech intensity and rate. Its dysfunction may explain the limited generalization observed in therapies such as LSVT. These therapies may work by temporarily shifting control to cortical areas, requiring patients to consciously monitor their speech. However, once the behavior becomes more automatic and control returns to the basal ganglia, the dysfunction re-emerges—leading patients to revert to their pre-therapy speech patterns. A similar phenomenon may be occurring in stuttering. When speech is de-automated, it often gets better for a while—but once it becomes automatic again, the patient relies on the same impaired circuit.*

Hélène Lœvenbruck (moderator/discussant)

Issue three concerns short-term adaptation, and how it relates to speech modes or speech styles.

We all know that speech is adaptive—and we've already talked a lot today about the kinds of adaptation that speakers do, some of which are production-based. There is what we are calling optimization of effort or cost on the part of the speaker. There is also adaptation to the listener, to the audience. This was elegantly articulated by Björn Lindblom, quite some time ago. In his "Hyper- and Hypo-articulation Theory"[10], laid out the various constraints that speakers have to deal with during communication. On the production side, he listed (i) physiological factors, which he described as involuntary and related to emotion and disease and (ii) cognitive factors, including situations where we speak to ourselves, not just to others. Today, we have focused mostly on speech directed at others, but self-directed speech (whether overt or covert) is also adaptive and can range from propositional to automatic. On the reception side, Lindblom highlighted social and communicative factors related to the communication channel, the listener or audience, the situation, the environment, and the degree of formality.

The adaptations to these various constraints influence speech along multiple dimensions. They affect intensity—from loud speech, like the examples Antje showed us today, to whispered speech, and even fully covert speech, internally produced speech. They also lead to variation in clarity: from carefully articulated, clear speech, to casual, reduced, or even highly condensed speech. There is variation in rate as well—from slow to fast speech. Prosody, of course, is also flexible, with speech ranging from highly melodic and intonational to flat and monotone. And finally, gestures are impacted. speech may be accompanied by expressive manual or facial gestures, by more subtle movements, or even by no visible gestures at all.

Now, related to speech adaptability, one of the questions that came up from you was: How are different speech modes encoded or parameterized for production? We have already touched on this briefly — mentioning that producing different speech modes likely involves adjusting motor plans rather than storing entirely separate plans for each speech mode. The fact that many of these adaptations are voluntary further suggests that we have active control over these adjustments. But this raises a deeper question: what exactly do we mean by *control* in this context?

We've already explored a few definitions of control, but now let's explore it introspectively. We can assume that what we have access to is what we monitor, what we can control. So let's try a quick experiment together—silently. Read this sentence silently:

"And I think to myself, what a wonderful world."

Now, play with it in your head. Can you add intonation? Can you sing it silently? Can you shout it—slowly? Can you shout it fast? Can you whisper it in your head? Can you imagine actually hearing Louis Armstrong sing it? This simple experiment illustrates that even inner speech can—at least for some of us—vary in intensity, clarity, rate, intonation, and even vocal quality. And crucially, it shows that we have a degree of voluntary control over these speech variations.

So how do we parameterize all of this? The fact that we can access and manipulate these different aspects and qualities of speech covertly suggests that we can actively monitor speech. That brings us back to the way we control overt speech. To introduce this, I will build on the consensual framework that Frank Guenther so clearly summarized and explained during this morning's session, namely what has been termed

[10] Lindlom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. *Speech Production and Speech Modeling. Kluwer, Dordrecht*, 403-439.

the ‘Standard Model of Word-form Encoding’, as presented in Levelt, Roelofs & Meyer (1999)[11]. I'll be using some of the terms Frank introduced, which we've generally agreed upon in our discussions today.

When we produce speech, we start with semantic content that is first conceptualized—this is the conceptualization phase. Production then proceeds through a formulation or encoding phase, followed by motor planning, motor programming, and finally, motor execution.

We know that speakers are capable of error correction, and can monitor for errors even before utterances are fully produced. This observation has led to the suggestion that monitoring is not solely based on external feedback, but also relies on internal feedback.

Levelt, Meyer and the Nijmegen team previously proposed that some form of inner speech—an internal signal -- is accessible before actual speech production, before motor execution. In many speech motor control models, this internal signal takes the form of a sensory prediction, essentially a simulation of the sensory outcome of the current plan, based on an efference copy of the motor commands. Crucially, this prediction can be decoded and analysed before execution takes place. It is assumed that speakers can compare the parsed prediction with the initial semantic content or the planned utterance. Interestingly, the internal signal, the prediction, is likely in a sensory format, which means it constitutes a form of inner speech. It is an inner voice that can be attended to and monitored. Covert speech, or inner speech, can therefore be understood as an exaptation, a by-product of overt speech control. The sensory prediction can be exapted. It can be used to speak internally without producing audible speech. If motor execution is halted, we are left with this this internal signal —a form of speech entirely contained within the mind.

In the predictive control model we developed in Grenoble, together with Marion Dohen, Maëva Garnier, Pascal Perrier, from GIPSA-lab, as well as Monica Baciu, Romain Grandchamp, Marcela Perrone-Bertolotti from LPNC, we proposed that the outputs at successive stages—preverbal message, phonological plan, and phonetic plan—can be understood as various formats of inner speech (ConDialInt model[12]). Depending on where the speech production process is inhibited, different formats of inner speech can be monitored. A late interruption, after sensory prediction, results in fully expanded inner speech, the familiar “little voice in the head”. An early interruption, after conceptualisation, yields the preverbal message, a fully condensed form. At intermediate stages, progressively less condensed and more expanded forms may be accessible. This accounts for the condensation dimension of inner speech, widely discussed in the literature (see [13]). The model accounts for another important dimension of inner speech: dialogality. The simulator that transforms the efference copy of motor commands into a sensory prediction— an internal model in some frameworks— can generate multiple voices. This is clear when we internally imitate someone’s voice, such as Louis Armstrong’s. The ability to imitate voices internally, reflects the simulator’s is adaptibility to multiple vocal patterns.

[11] Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and brain sciences, 22(1), 1-38.

[12] Grandchamp, R., Rapin, L., Perrone-Bertolotti, M., Pichat, C., Haldin, C., Cousin, E., Lachaux, J.-P., Dohen, M., Perrier, P., Garnier, M., Baciu, M. & Lœvenbruck, H. (2019). The ConDialInt model: Condensation, dialogality, and intentionality dimensions of inner speech within a hierarchical predictive control framework. *Frontiers in Psychology*, *10*, 2019.

[13] Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, *141*(5), 931.

Sticking to this consensual framework, several questions arise: where does adaptation occur? Are phonological plans, phonetic plans, or motor plans adjusted? At which stages are internal models actually fine-tuned? Or do we, in fact, rely on distinct conceptualisers, formulators, motor planners/programmers, and simulators? Does adaptation occur simultaneously across all stages, as suggested by Ben Parrell?

Another point we haven't addressed today is parsing of internal signals —but we will probably revisit it in issue four. We have a parsing mechanism that converts the sensory prediction, originally in a sensory format, into a representation that is more semantic and thus comparable to the initial goal or semantic content. Is this parser itself tuned? It can be decomposed into several decoding units (I won't go into that now), but do we fine-tune those as well? These are some of the questions we would like to address now.

Frank Guenther

Can you clarify a bit what you mean by tuning in this case?

Helene Lœvenbruck

I mean adapting. For instance, when we adapt to a situation, because we want to speak more clearly, faster, or add specific prosodic features. Is it only at the motor programming stage that we adapt?

Frank Guenther

You mean a rapid change, basically?

Helene Lœvenbruck

Yes, I mean online, short-term adaptation.

Frank Guenther

In the DIVA model, the way we implement clear speech is by shrinking the target region. Each dimension in the planning space corresponds to a region rather than a single point. When we shrink those regions, the model effectively produces hyper articulated speech. To me, that adjustment is likely planned rather than at the motor execution stage, because it involves changing the targets. According to our definition of motor control, that would fall under planning rather than control.

That said, I do think that adjustments also occur at the motor control level—things like rate and probably intensity. But overall, my suspicion is that across nearly all of these stages, some form of feedback is used to monitor and potentially tune productions.

Ben Parrell

Yes, I think clear speech is a good example, because all of these adjustments aren't typically parameterized in a single dimension. Shrinking the target region in the DIVA model, for instance, results in articulatory hyperarticulation, but it doesn't capture changes in duration, intensity, or pitch that also occur simultaneously in clear speech—and that seem to be not independently controlled in the model. So is it an arbitrary linkage at this point?

Frank Guenther

You shrink the targets and you extend the duration.

Ben Parrell

It seems like there is a separate dimension that we outlined here—one that allows factors like arousal to influence multiple levels of the production process. For instance, clear speech has specific consequences, but arousal might lead to a different pattern: speaking louder and faster, rather than louder and slower. Arousal may shorten reaction times, too. These effects appear to be linked. So, I think there is another dimension at play—something that isn't explicitly represented here, but that can shape behavior across multiple stages of the system.

Antje Mefferd

I would like to add to that as it relates to my talk. When an articulatory movement requires a certain amplitude or distance, given there are no other constraints, the speaker will likely execute the movement at a speed that requires the least amount of effort. But if there is an additional constraint—for example a durational constraint—then the speaker has to increase speed and exert more physical effort to complete the task (e.g., [14].). So, the relationship between speed and duration really depends on how much effort the speaker is willing to exert or the task demands.

Doris Mücke

And picking up on this idea of multidimensionality—together with Lena Pagel[15], Simon Roessig, and Márton Sóskuthy, we investigated prosodic prominence in a multimodal data set including articulation, f0 and visual head-movements[16]. We know that placing an accent in habitual speech involves modifications in various dimensions. What we observed in loud speech is a similar behavior, but the degrees of adjustments of the individual phonetic parameters differ. It seems that the relative importance of the channel varies depending on the speaking style. In loud speech, articulation is more modulated, while f0 is less modulated. At the same time, visual cues—like co-speech gestures—become more prominent in loud speech.

I think this is quite interesting: not only are different channels involved, but the relationship between those channels also changes depending on the speaking style. And to me, that suggests these effects are planned in advance—not something that just happens during execution by chance.

Ben Parrell

Yes, and I think there's also this ongoing re-evaluation of what we actually mean by *clear speech*. Speaking in a loud environment, speaking to someone who's hard of hearing, or speaking to a non-native speaker—all of those contexts fall under what we often call "clear speech," but they actually lead to different adaptations. I think the different communicative contexts have different effects on what we call clear speech.

[14] Nelson, W.L., (1983). Physical principles of economies of skilled movements. Biological Cybernetics, 46, 135-147

[15] Pagel, Lena, Simon Roessig & Doris Mücke. (2024). The encoding of prominence relations in supra-laryngeal articulation across speaking styles. *Journal of Laboratory Phonology* 15(1), pp.1-55. DOI: https://doi.org/10.16995/labphon.10900

[16] Pagel, Lena, Sóskuthy, Márton, Roessig, Simon, & Doris Mücke (2023). A kinematic analysis of visual prosody: Head movements in habitual and loud speech. Talk at *International Congress of Phonetic Sciences (ICPhS)*, 7-11 August, Prague, Czech Republic, p. 4130–4134. DOI: 10.5281/zenodo.10299230.

And to Doris’s point, it’s not just as though there is a single “clarity” dimension—it’s that certain aspects are emphasized more or less depending on the context. These things are all interrelated and highly contextual. The key question becomes: what are the actual constraints on communication that I'm trying to maximize? It is a complex question, but I think that's the right way to go about it.

Hélène Lœvenbruck

And when we consider brain regions, it’s already challenging to pinpoint exactly which areas are involved in each of these stages. Adding the cognitive control that speakers seem to exert over variation and adaptation makes the picture even more complex. So I was wondering—Frank, in your model, where would you place the online parsing that speakers perform? And do you think this parsing mechanism should itself adapt when speakers adjust their speech?

Frank Guenther

So by *parsing*, you mean the process of taking the acoustic signal and breaking it down into things like words and other linguistic units?

Yes, these processes are happening in parallel. For the planning part specifically, I think premotor cortex is likely where some of these adjustments are made. And I think these are testable hypotheses. For instance, has anyone ever scanned someone while they’re faking an accent? That could be really informative—it might show us where in the brain this kind of adaptive processing is happening. My guess is that you would see activity in the premotor cortex. But depending on what exactly you’re manipulating, I suspect different brain regions will be involved.

Hélène Lœvenbruck

And Doris, perhaps you could add something about prosody—where do you think prosodic adaptations take place in the system?

Doris Mücke

It’s a very good question—where does prosody come into play? And as we heard earlier this morning, where is the metrical structure inserted? This whole chunking process is really complex and difficult to localize.

Frank Guenther

I just wanted to add to that—I think there’s more than one aspect of prosody. Emotional prosody, for example, is probably very different. But linguistic prosody seems to me like it has to come in at least at the planning level, or even earlier, because it is part of the meaning—it is part of the linguistic message. So I’d expect it to involve the left inferior frontal areas and the premotor cortex. Somewhere in that range is likely where it is being inserted.

Marina Laganaro

I think we can agree that tuning or adjustments happen both at planning and programming. But now imagine someone who regularly uses loud speech much more than I do—for instance, a teacher. Or someone who uses whispered speech more often—for instance, someone who works all day in a library.

Do you think, over time, they end up storing speech plans or motor maps that are specific to these different speech modes?

Ben Parrell

I wouldn't think they store individual words as, say, a whispered version or a loud version. But I do think your earlier point about whispered speech having longer reaction times is relevant. My guess is that if someone is used to whispering regularly, they probably wouldn’t show that delay. Whispering is an unusual coordination pattern for most people, and that takes more time to execute.

So I think it's really about habitual learning—like Pascal mentioned earlier. We get used to certain speech behaviors, and we get better at them simply because we do them more often.

Marina Laganaro

So is it still tuning, perhaps faster tuning?

Ben Parrell

I think so. It's still freely combinable.

I think what’s interesting is that these seem to be global settings that interact with local motor plans. So it’s not entirely clear to me how that parameterization fits in. But I think, as Frank (Guenther) suggested, they probably operate at multiple levels—both planning and programming. And like with many things, when you’re used to it, it becomes easy; when you’re not, it’s harder.

Antje Mefferd

I just want to add that this is actually my conundrum when using a loud speech approach in therapy with patients with Parkinson’s disease. We're practicing loud speech, and there is data that suggests that it increases effort. One rationale for the use of loud speech is that it can help recalibrate the speech motor system and that it can help patients use more effort again. But in practice, I’ve rarely seen patients execute loud speech independently unless they were explicitly cued.

And if we consider that Parkinson’s disease is associated with a pathology of the basal ganglia—and the basal ganglia are involved in motor learning—then I question that these patients can truly learn to change their speech. Thus, although there is evidence that loud speech can target their problem, the reality is that therapy doesn’t always carryover into every day conversations. That, I think, is a major challenge with this intervention.

Marina Laganaro

But then it seems to us that it involves more ‘planning 2’, what we call programming.

Frank Guenther

I think we haven’t talked much about subcortical structures, but the basal ganglia are involved at all these levels—they receive input from the prefrontal cortex, premotor cortex, motor cortex. They seem to play a role in regulating intensity or speed; there is a signal coming from the basal ganglia, or it is part of the controller for that.

What I think is happening in these therapies, in my opinion, is that you’re forcing the cortex to take over the job by making patients consciously think about what they’re doing. But as soon as the behavior becomes automated again and shifts back to the basal ganglia, the basal ganglia doesn’t work properly. That’s why patients revert.

And I don’t think this is unique to Parkinson’s therapy. I think we see something similar in stuttering. When you de-automate speech, it often gets better for a while—but once it becomes automatic again, you’re relying on the same impaired circuit.

Hélène Lœvenbruck

There was also a question raised by the audience when we first conducted the survey: how can we ensure that individual variability is properly taken into account? Importantly, this question is not limited to pathological speech—it also applies to typical speech.

Then closely related is the issue of whispered speech: how should we account for whispering in conditions like apraxia of speech or dysarthria, and how does that compare to voluntary whispering in healthy individuals?

Ben Parrell

I think we all rely on internal predictions to a different degree. We do have these internal predictions, but some of us are more reliant on them, while others rely more on actual sensory feedback to monitor how loud or how fast we're speaking.

In altered auditory feedback studies, we see a large range of variability in how much people respond. Some have suggested this might be due to differences in auditory acuity—and there's some evidence for that—but I think there is a lot more to it. It could also be that some people just don’t care as much about how they sound, and others do. Some might say, "Well, my predictions are really good, I’ve learned that," and so they rely on those rather than on monitoring the outcome—unless there's a breakdown in communication that forces them to recalibrate. This becomes kind of a theory of mind issue, in a way.

Frank Guenther

And there are experiments showing that people differ in how much they rely on auditory versus somatosensory feedback—like the work by Daniel Lametti[17], for example. His studies show that these individual preferences or weights on different feedback modalities really do vary, which probably plays into this whole spectrum of monitoring strategies.

Ben Parrell

Right, but those are about relying on different sensory signals versus relying on prediction over all sensory signals.

Frank Guenther

[17] Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *32*(27), 9351–9358. https://doi.org/10.1523/JNEUROSCI.0404-12.2012

Yes, I guess what I’m saying is that some people are better at making predictions and some people rely more heavily on their auditory feedback. It’s the same basic idea.

Wolfram Ziegler

There was a question about whispering in apraxia of speech. I think whispering is actually quite different from the other speech modalities we have been talking about, because it uses a different laryngeal gesture. From a gestural point of view, that means it is a different organization of syllables. And the practicing speakers may be very sensitive to the low frequency of the laryngeal gestures we use in whispering.

Ben Parrell

I would add that it's not only the presence of a gesture, but it actually is the removal of typical opening and closing gestures that you'd have in normal speech too, so it does seem really different.

Pascal Perrier

Just to add to what you said—not only do we have to consider how individuals differ in how they take into account sensory information, but also the differences in the vocal tract itself. For example, many differences in control are simply due to the fact that our palatal shape is not the same. So, the sensitivity of the acoustic output used to change the movement is simply different as well. That, I think, is the first way to take into account inter-speaker variability, even among healthy individuals.

Frank Guenther

Along those lines, way back when we did make speaker-specific vocal tracts and used them to control movements, which are widely variable across individuals. And just from the different physics of each vocal tract and the way the DIVA model learns, the model ended up imitating the same speech gestures, even though the actual gestures differed from person to person. The model reproduced what each speaker did, which really highlights that it was the vocal tract itself that was really changing the way they speak.

Louis Goldstein

One small observation about whispering that I’ve always found intriguing—though I’m not aware of any modern data on this—is based on the conventional wisdom going back to J. C. Catford[18]. The idea was that during whispering, typical speakers replace the state of the larynx for voicing with whisper, but the gestures for abduction states for voiceless sounds remain unchanged. So the adjustment is said to apply only to the voice during the whisper stage. I don't know if anyone has really examined that carefully, and whether that would be the same for when you find whispering in typical speech.

Doris Mücke

I have a question for Hélène. If inner speech and imagined voices remain intact even when overt speech is impaired, what does that imply for the notion of motor programming? For instance, if a speaker shows disordered speech, does that also affect their inner voice?

Hélène Lœvenbruck

[18] Catford, J. C. (1964). Phonation types: the classification of some laryngeal components of speech production', in Abercrombie, D., et al. (Eds.). *In honour of Daniel Jones*, 26-37. London: Longmans.

Excellent question. In people with acquired non-fluent aphasia—for example, after a stroke—it is of course, very difficult to rely on questionnaires, because these patients have language and speech production deficits. But among those we have been able to interview, many report that their internal speech is not impaired and feels similar to their speech prior the stroke, their “previous voice”. That said, not all aphasia patients maintain inner speech. In some cases, inner speech itself is affected—for instance, in individuals with severe anomia or lexical access difficulties. This suggests that brain lesions can disrupt different stages of speech production. When the damage affects only the later stages, we can reasonably suppose that the sensory prediction—what we call the inner voice—might still be experienced by the patient.

Another interesting case involves individuals with cerebral palsy. For example, Frank, when working with patients—have you asked whether they could hear their inner speech?

Frank Guenther

We did. We had locked-in syndrome from a brainstem stroke, but he could still speak and sing in his head. And when he was actively trying to speak, everything seemed to happen in his head, just like before, except no sound would come out. It was as if everything but the actual vocal output was still intact.

Hélène Lœvenbruck

So it is likely that the impairment affected the final stages of speech production—motor execution, and possibly motor programming.

# Round table 4: Learning, changing, adapting speech

*Wolfram Ziegler (moderator), Louis Goldstein, Frank Guenther, Hélène Lœvenbruck, Ben Parrell, Pascal Perrier*

## *Summary*

*Round table 4 addressed the mechanisms of learning and change in speech across the lifespan and the related issues of adaptation as well as the clinical perspectives and the role of learning and adaptation in speech models.*

***Mechanisms of change: Learning, reinforcement, settling, variability*** *Multiple forms of learning—procedural and sensorimotor—operate on different timescales. It was discussed as a mechanism not only in infancy but also in mature speech systems. While procedural learning in the basal ganglia allows for rapid adjustments—such as after a gross sequencing error—slower, cerebellum-driven sensorimotor learning fine-tunes motor control over time. Reinforcement may guide both early speech development and ongoing refinement through feedback. Reinforcement isn't just about correction; it may also help shape speech when communicative feedback is ambiguous or delayed, as in interactions between non-native speakers.*

***Adaptation and interaction: Phonetic adaptation, accommodation, social interaction*** *The discussion ranged from short-term phonetic adaptations to long-term change across the lifespan and even historical language change. Phonetic targets and motor plans are plastic, and different types of learning happens at different time scales. Phonetic adaptation is mediated by more than sensory-motor alignment; it also involves social and cognitive goals. Speakers adapt to the phonetic patterns of others across time. Accommodation is not symmetrical: the more variable speaker tends to shift toward the more stable one.*

***Clinical perspectives on adaptation and accommodation*** *Individuals with cerebral palsy who use speech synthesizers still have inner speech and comprehension, suggesting the presence of high-level planning even in the absence of motor programming and motor execution. The cerebellum supports feedforward prediction; adaptation and accommodation is compromised in cerebellar disorders but preserved in others. Patients with basal ganglia disorder may even show hyper adaptability.*

***Speech models*** *Participants largely agreed that current speech production models are simplified, often for conceptual or mathematical tractability. There was strong consensus that models should eventually incorporate interactivity, variability, and learning over time. Multiple speakers pointed out promising alternatives—like Bayesian approaches, or dynamical field models—that can capture interaction effects. They allow continuous adjustment of target representations in response to input variability.*

Wolfram Ziegler (moderator/discussant)

The issue I'd like to raise concerns the fact that language is, for the most part, used interactively. So the question is: how does the interactive nature of language use impact speech production and how should it inform our models of speech production?

I think that language learning, especially in childhood, is completely interactive. Yet the models we typically use are entirely non-interactive. They do not include speaker–listener interaction. From a technical standpoint, model training is treated as a fluid process—but once trained, the application of the model is essentially frozen. As a result, the auditory and motor target representations assumed in these models are unaffected by speech input from others. That means that they will remain unchanged throughout life.

In reality, we know that speakers are constantly surrounded by other speakers, and that there is a great deal of variation within any language community. So the question is: to what extent are we, as speakers, influenced by the variability in the language surrounding us?

Take, for instance, someone raised in Marseille who later moves to Normandy. Over time, that speaker may adapt to the local dialect. This can be a slow process, but aspects of their speech may change—like their vowels or their consonants. So the question would be: once phonetic knowledge is acquired, is it really sealed off from phonetic variation in our language community?

We know that it isn't. There is abundant literature showing that phonetic changes over the lifespan. One recent and compelling example comes from Jonathan Harrington and his group[19]. They studied a group of researchers who spent six months together in Antarctica. Over the course of their stay, the researchers began to develop the early stages of a common accent and converged in their vowel articulation.

There are several brain imaging studies—such as those by Stephens and colleagues[20]—that support the idea that we adapt very fast to other speakers during speech. See [21] for a recent overview and discussion.

We ran a series of experiments with patients who had brain lesions. In patients with cortical strokes, we found that even large lesions in the right hemisphere did not influence the adaptation behavior in these patients. Similarly, patients with lesions in the *anterior* part of the left hemisphere also adapted very much to variation in other speakers. In contrast, patients with lesions in the *posterior* left hemisphere did not adapt to any variation in a model speaker ([18]).

We also ran a second set of studies with patients diagnosed with Parkinson's disease and cerebellar degeneration. Here, we saw that patients with cerebellar degeneration in SCA6 had severely compromised adaptation. Interestingly, patients with basal ganglia disorders, such as Parkinson's disease, showed preserved adaptation, and in some cases even hyperadaptivity ([18]).

So, perhaps we could propose an expansion of this model. Take, for instance, a rough sketch of the DIVA model, which includes initiation, motor planning, and articulation in the feedforward pathway on the left side, and the feedback loop on the right. I would particularly emphasize the forward modeling part, where the cerebellum and the left posterior superior temporal gyrus are deeply involved.

A prediction from the model proposed by Pickering and Garrod[22] among others, is that when we listen to someone else speak, we covertly imitate or emulate their speech on a motor basis. In this view, while we use the incoming signal primarily for auditory comprehension, we simultaneously generate a motor-based prediction for the auditory outcome of their speech. This predicted outcome is not identical, but may overlap with our own forward model. If a consistent difference emerges between the other speaker's

[19] Harrington, J., Gubian, M., Stevens, M., & Schiel, F. (2019). Phonetic change in an Antarctic winter. *The Journal of the Acoustical Society of America*, *146*(5), 3327. https://doi.org/10.1121/1.5130709

[20] Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. Proceedings of the National Academy of Sciences, 107(32), 14425-14430.

[21] Ziegler,W. & Aichert, I. (2025). Phonetic adaptation and rhythmic entrainment in interactive language use: Neural mechanisms and evidence from individuals with neurological disorders. In: Meyer, L. & Strauss, A. (eds.), Rhythms of Speech and Language: Culture, Cognition, and the Brain. Cambridge University Press

[22] Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and brain sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

speech and our own, then there may be a slow change in our ownauditory targets regions—and potentially in our motor plans as well.

Based on our data, the cerebellum appears to be the key structure supporting this kind of prediction and feedforward adaptation. And we did not observe any adaptation deficits in patients with basal ganglia disorders or apraxia of speech.

This is, of course, only a brief sketch, but it outlines what we might expect from such a framework.

Ben Parrell

I think this connects to a broader point I've been wanting to make during this conference: there isn't a single, definitive speech model. We all develop these models to examine different things. Take DIVA, for example—you have developed it over a long time, and it has changed a lot over the years as you've asked more questions and pushed the model in different directions.

One thing we don't often make clear is the distinction between what is theoretically important in a model and what's simply included to make the model mathematically tractable or publishable. So when we say that a model has fixed targets that are invariant, that shouldn't necessarily be taken as a strong opinion. It's often just a way to isolate and test a specific component—like motor control.

In reality, it's obvious that we continually adapt our production to the people around us in the way that you mentioned. Another way to conceptualize this is through exemplar theory—constantly hearing distributions over different acoustic parameters. These could serve as the input to something like a Bayesian model of speech production, in which we select the most optimal thing to produce what we think is associated with a particular linguistic category.

I am totally in agreement with this whole framework.

Wolfram Ziegler

I'm bringing this up because I see these processes as the core of something much larger—namely, accent drift, language drift, and, over the long term, the diachronic changes we observe in languages. These interactions are the seedbed of all of that.

Frank Guenther

Just to add to that—for this particular case, I'm pretty familiar with why we made the choices we did. And in fact, it really was for simplification. But I do think the auditory targets probably do change over time. You can adjust the motor plan quickly but the targets themselves, I suspect, are more like synaptic-level representations that take a long time to evolve. So surely, those targets aren't fixed from the beginning.

On the developmental side, our model assumes perfect auditory targets from the start—but real children don't even hear all the right things the first few times. They must be updating their model as well over time. So, yes, the current setup was made to keep simulations tractable, but the model is definitely amenable to these changes—if someone wanted to incorporate them based on empirical data.

Louis Goldstein

There have been a number of models that develop this theme using dynamical fields—where a speaker's target isn't represented in a particular set of sensory coordinates, but rather as activation along a continuum. That continuum represents narrow ranges of sensory or motor values, and in an interactive environment, all kinds of input change the resting activation levels and effectively shift where the mode of the distribution lies.

So these models exist, and like what Frank and Ben said, there are such models that make some good predictions. In our own work ([23]), for example, we found that when speakers accommodate to one another, it's typically the speaker whose distribution of values is more variable *before* the interaction who shifts toward the other speakers. The accommodation is asymmetric and predictable, based on the variability in their baseline distribution. Theoretically, this relates to the relative activation within the width of the range of activations that the neural field corresponds to.

So, while models like the ones we've discussed here today may not include those dynamics—often for the sake of simplicity—there are models that do focus on this aspect. And of course, when modeling, we often have to study pieces of the system independently rather than trying to simulate everything all at once.

Ben Parrell

I think this ties into the earlier conversation we were having about motor planning. If your underlying representation is a distribution, then at some point during the planning process, that distribution settles into a parameter that gets passed on to the articulatory system. After all, you don't produce a distribution; you produce one action.

I think that the process of settling—or selecting a point within the distribution—is part of planning. We go from a distribution that is affected by both our own speech history and what we are currently hearing in our environment, to the one thing we are actually going to produce every time. And each time we produce a sound, what we actually do may be slightly different, depending on stochastic variation in that dynamical field.

Wolfram Ziegler

There are also shorter-term adaptations that happen during dyadic interaction. In turn taking, for instance, you predict when the other speaker will end their turn and begin planning your response while still processing auditorily what they are saying.

Ben Parrell

I think that all of our current models in the production world are really just that—production models. We might include perception to alter the targets, but they won't be models of conversation where linguistic planning happens dynamically in response to incoming speech. There is some really interesting data from Greg Castellucci[24], that looks at the timing of planning during perception. But so far, these insights haven't been integrated into articulatory models.

[23] Lee, Y., L. Goldstein, B. Parrell, D. Byrd. (2021). Who converges? Variation reveals individual speaker adaptability. Speech Communication. 131:23-34. https://doi.org/10.1016/j.specom.2021.05.001

[24] Castellucci, G. A., Kovach, C. K., Howard, M. A., 3rd, Greenlee, J. D. W., & Long, M. A. (2022). A speech planning network for interactive language use. *Nature*, *602*(7895), 117–122. https://doi.org/10.1038/s41586-021-04270-z

Pascal Perrier

I’d like to come back to the issue of generating motor-based predictions for other people’s speech. I don’t see how we could infer another person’s motor plans. When we form our own motor plans, we have a complete sensory input at all the levels up to the auditory feedback. Whereas when we listen to someone else, we only have access to their auditory signal, and maybe some visual cues like lip movement. But their internal motor plans aren’t accessible to us.

For me, the important point is that if we have internal models, they can only mimic what we’ve learned ourselves. They can’t reproduce what’s inside another person’s system. So the only output we can rely on is auditory feedback. We certainly have a prediction and try to adapt our motor plans accordingly to reach the auditory goals. There is a slow adaptation of our own auditory targets and that might, in turn, change our somatosensory goals accordingly. But there are never changes in the biosensory of a person.

Wolfram Ziegler

The idea of emulating other speakers isn’t mine—it comes from Pickering and Garrod. And I agree that we cannot completely jump out from our target regions. We stay within the boundaries of what we’ve learned to produce. But we do gradually shift toward the speech patterns we’re exposed to.

Frank Guenther

Another way to think about this is that our auditory space adjusts depending on who we are speaking with—much like our visual perception changes in a dark room or a large versus small space. So, there are adjustable aspects of auditory space and that may in fact change our productions.

Louis Goldstein

When I imagine speaking like a particular person, it's not just an auditory impression—I can actually see and hear them in my mind. And interestingly, in my mental image, they don’t move their jaw much when they talk, so I imitate that too. So it’s not purely auditory; auditory information is part of the source, but I can actually see the speaker in my mind.

Ben Parrell

Another important point to keep in mind is that there's a whole literature on phonetic accommodation showing that it is highly mediated by social dynamics. It's not just an automatic recalibration of sensory and motor systems—it's also about our interactions with other people. Whether we want to affiliate with someone, how we relate to them, whether we want to sound like them or not—all of that plays a role. So adaptation in speech isn't purely physiological; it is also social.

Wolfram Ziegler

So yes, actually, the cerebellum is also known to take part in social interactions. Patients with cerebellar disorders often have difficulties with these aspects of social interaction as well.

Hélène Lœvenbruck

Thank you, Wolfram, for summarizing the Pickering and Garrod account. I also really appreciated Pascal’s point, because the format of the motor plan is crucial for understanding how individuals with cerebral palsy, for example, can interact. Individuals with congenital cerebral palsy who use a speech synthesizer to communicate, still possess inner speech. Some have an inner voice, which they encode by typing on a keyboard to generate synthesized speech. So we can assume that some level of speech planning occurs, because they know which speech units they intend to produce. But they don't have access to motor execution, and likely not to motor programming either. Yet they can fully decode incoming speech and understand other people's speech. If, as some accounts suggest, understanding others involves simulating their motor plans, then we need to clarify what a motor plan really is in this context—because clearly, these individuals are able to perceive and understand speech without engaging in motor programming/execution themselves.

Audience

I think what you mentioned—the importance of stochasticity or variability—is really important here. We've been talking about linguistic goals, but we've also hinted that there are many other goals involved in speech. Communication only works if there's some flexibility—some slough in the system. You need to have the ability to index all the other things that language indexes in order to communicate. If you hit the perfect target every single time, you actually can't do that. In fact, I'd argue that this stochasticity is essential for the system to work. It is not about an ideal target.

Ben Parrell

Yes, variability is not that, I agree.

Audience

I think, to some extent, we can reformulate speech production as a reinforcement learning problem. After all, our ultimate goal is to be understood. When it comes to learning, how does the model weigh a backchannel such as "Was my utterance understood? To which extent?" Of course, we see short-term adaptation in interactions. Beyond that, there may also be long-term learning that emerges from this speaker–listener interaction, perhaps even shaping the development of speech production in the first place.

Ben Parrell

I think there are two questions here. The first concerns the role of reinforcement in development. Some have suggested that early motor programs are reinforcement-based learning rather than sensory motor learning. You produce a sound, your mom smiles, and you learn: *that's good, do it again.*

The second question is: once we have a mature system, how does reinforcement push us? That relates to what Pascal brought up earlier—we have a well-developed motor repertoire, practiced routines, and while we can maybe push to the boundaries of that, pushing beyond is really challenging.

You end up with situations where long-term misunderstandings persist between people. Imagine two people are communicating in a shared language that neither of them speaks fluently. That's a really challenging situation because you don't have the right ability to shape your motor system to be understood and reach the sort of common targets. How reinforcement interacts with sensory motor learning is an interesting question.

Frank Guenther

And I think it's important to keep in mind that multiple types of learning are happening simultaneously. There is procedural learning based on reinforcement that the basal ganglia learn fairly quickly, within a few trials. Meanwhile, tuning the motor system through sensory–motor learning in the cerebellum is a much slower process.

Say you hear yourself make a mistake, like a gross sequencing error. You might start over or enter a kind of tinkering mode, where procedural learning quickly readjusts your output. Then, you probably fine-tune that over time. There are definitely different timescales of learning at work in parallel.

Wolfram Ziegler

One question that comes to mind in this context is that all of these—motor targets, sensory targets, auditory targets—are essentially memory systems. So what happens if someone only ever speaks to themselves? Imagine a scenario like a perturbation experiment, or a hermit living in a cave, entirely isolated from interaction with others. Would this person perhaps lose these memories, or would they remain stable?

Frank Guenther

I suspect that they will continue to respond to small perturbations.

Ben Parrell

I think in terms of their motor execution, it would probably become problematic over time. After all, the purpose of the sensorimotor adaptation system isn't primarily to deal with the weird and external perturbations we create in the lab—it's there to constantly maintain our speech production accuracy. So if someone didn't speak for years, I would expect some neural drift. Their motor programs just wouldn't be as well-tuned anymore, and if they tried to speak it would not be perfect. But if that person would be speaking in isolation, just not to someone else, then I think it would actually be fine.