情報科学研究科

School of Information Science

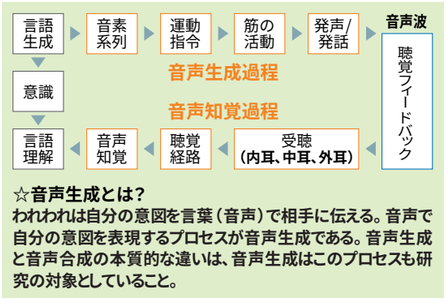
党研究室

機械の“ 声 ”は、ここまでヒトに近づく。

**芸術は自然を模倣する、というのはアリストテレスの言葉。
そこには自然の中にこそ、真理があるという概念がある。
今、先端技術でも機械が人間という自然を真似ようとしている。
党研究室では、人間の音声生成のメカニズムから、
新たな音声認識・合成の技術を生み出そうとしている。**

▼音声の研究の歴史。
～人間の発話メカニズムを真似るか、あるいは音声を信号として捉えるか。

**ヒトとヒトが交わす言葉、音声は、重要なコミュニケーションの媒体のひとつ。音声の研究の歴史は長い。17世紀には機械で人間の発話を真似ようという研究がさかんに行われていた。20世紀に入ると電子的にそれを実現しようとした。
ところが60～70年代には、それまでの人間のメカニズムを真似ようという流れから一転、コンピュータで音声を信号として認識、合成しようとする研究が主流になった。コンピュータの性能が向上するにつれ、音声信号処理の分野も進歩した。現在、雑音のない環境で朗読された言葉を認識するなら、ほぼ実用化できると言える。
それでも感情が込められたコトバを認識することは難しい。従来の研究では人間の音声生成のメカニズムが考慮されてないから、と党先生は言う。
党先生の研究のコンセプトは、もう一度、人間のメカニズムを見つめ直すこと。音声の裏には、個性や感情などの情報が含まれている。ヒトはどういうメカニズムでこうした情報を音声の中に含めているのだろうか。それを探っていくことで、個性があって自然な音声合成の技術や、新たな音声認識の技術を確立することを目指す。**



☆音声生成とは？
われわれは自分の意図を言葉（音声）で相手に伝える。音声で自分の意図を表現するプロセスが音声生成である。音声生成と音声合成の本質的な違いは、音声生成はこのプロセスも研究の対象としていること。

▼音声生成のメカニズムと、音声の個性。
～人間はどうやって音声を発しているのか。どうして音声に個性があるのか？

「人間はすべての発話器官を調和させています。たとえば何か食べていて舌が動かない場合でも、別の部分を動かして補正して、聞き取ることのできる言葉を生成している。全体的に一つのシステムとして、自分の出す声を聴きながら喋っているんです。また会話しているときも実は相手のはなしの、一つひとつのワードについては100%分かっているわけではないが、全体の意味はほぼ完璧に理解できている。生成と聴覚は表裏一体なんですね。発話器官の一部分だけを見ていても、どの筋肉が関与してこんなことが可能なのかは分かりません」。

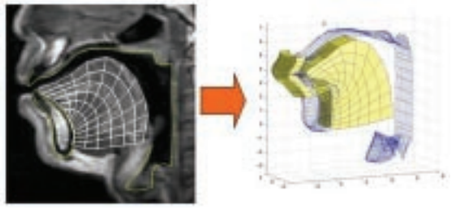
音声は、脳の指令によって声帯や舌、口唇など発話発生器官の筋肉が駆動されることで生成される。もちろん発声発話器官の大きさや構造には個人差があるから、その差が個性として音声に表れる。

「音声の個性は二つの要素によって決まります。ひとつは性別など生まれつき備わった要素、もうひとつは方言や教育などの習慣として身についたもの。わたしたちは生まれつき備わった生理的な要素を対象にしています」。

人間がどうやって発話するのか、そのメカニズムを解明することが、個性のある音声の合成・認識や、様々な応用につながっていく。党研究室はその第一歩として人間の発話機構のモデルを構築している。

▼形状と動き。
人間のメカニズムに忠実なモデルをつくる。
～筋肉が動いて生成される音声。逆に音声から筋肉の動きを推定する。

党先生は、MRIで特定話者のデータを取り出し、生理学的データに基づいた忠実な発話モデルを作成している。これは人間が話すときの基本的な三次元の動作を再現できる。ヒトは舌や唇を動かして音声を生成する。モデルを使えばその逆で、音声から舌や唇の筋肉の動きを推定できるのである。

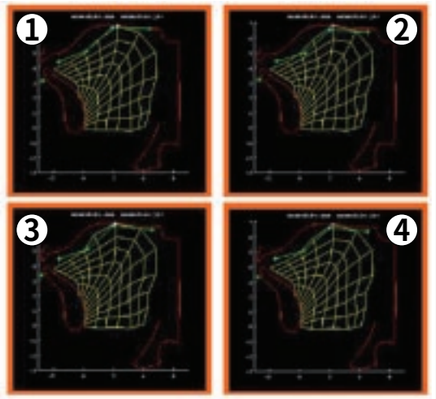


☆ヒトの発話器官はどんなカタチ？（形状を計測する）
- **正中矢状断面と1cm外側の傍矢状面のMR画像より抽出した輪郭により構成され、左右幅2cmの厚みをもつ。**
- **このモデル上に同話者のMRIデータに従い舌筋を配置する。**
- **声道壁に相当する口蓋、咽頭後壁および下顎の表面は左右幅3cmの剛体壁からなる。**

☆ヒトの発話器官はどう動くの？（動きを観測する）
- **発話意図に基づいた運動指令に従い、発話器官を駆動して声道形状の時間変化を生成する。**
- **有声音の場合には声門に周期的なパルスを発生させる。**
- **無声子音の場合には声道の狭めにノイズを生成する、あるいは閉鎖点に破裂音源を発生させる。**
- **声道形状の変化により音色を調整し、最終的に口唇または鼻孔から音声波を放射する。**

「たとえば“ おはよう ”と言ったときに、どんなふうに筋肉が動くのか。このモデルを用いれば分かるわけです。同じ言葉でも、怒ったように言うのと、感心したようにが言うのでは筋肉の動きは違います」。

異なる方法で発話した/ka/の動画像
（舌背の軌跡を太線で表示）



①がっかりした　②関心した　③疑った　④軽い相槌

これは逆推定、という技術である。どんなことに応用できるのだろうか。

「子どもは大人が話すのを耳で聴いて言葉を習得していきます。でも耳の不自由な人はこうしたオーディオフィードバックを受けることができません。そこで耳で聴く替わりに目で見えるかたちで確認するビジュアルフィードバックが有効だと思うんです。言葉を発すると自分の舌の動きが映像でフィードバックされて、先生の舌の動きと比べることができる、というような。英語を勉強するときも日本人が苦手なLとRの発音の違いを舌の動きの違いを見て覚えられるかもしれませんね」。

**従来のFEM技術だと、1秒の音声を生成するのに10000秒かかる。党先生はモデルの計算方法を改良することで、100倍のスピードアップに成功している。実用化を考えると、モデルの精度を上げることはもちろん、さらにリアルタイムにフィードバックできるスピードが欲しいという。
コンピュータグラフィックが進歩している今、本当に喋っているかのように口を動かすアニメーションもそれほど珍しくはない。しかし、そうしたCGと決定的に違うのは、あくまで人間の機構に忠実であるということ。
「なぜ人間の機構にこだわるかというと、このモデルは言語の病気、失語症などの原因の解明や治療法に役立つから。人間はブラックボックス。言語の病気の原因は、脳なのか、脳の指令を伝達する部分なのか、あるいは筋肉なのか、というのは分からないんです。人間と同じように動くモデルがあれば、これを確かめることができるはずです」。**