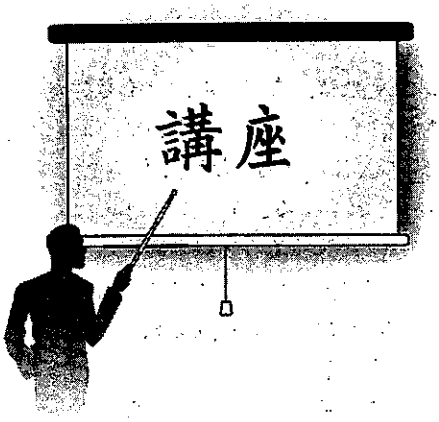

コーパスベース音声合成技術の動向[V・完]

——大規模音声コーパスによる音声合成——

Technology Trends in Corpus-based Speech Synthesis [V·Finish]: Developments in Corpus-based Speech Synthesis

Nick CAMPBELL

1. はじめに

近年，コーパスベース音声合成は，大規模音声データを基にし，高品質な声質と自然性の高い合成音声を可能にした．本稿では，次世代の音声合成に向けて，3点を提言する．具体的には，コーパスとは何か，音声合成の利用について，発話音声が表現する情報について述べる．現在音声合成は主に情報提供に使われている場合が多いが，今後は更に，人間の代りに使われるアプリケーションが増えるといえよう．

2. 波形接続型音声合成

音声合成の研究において実音声データから，音響的特徴分析とともに，韻律や発話タイミングなどの制御パラメータが検討されてきた．しかし，現在では，この音声データは合成研究の基礎データであるのみならず，合成時においても利用されている．つまり，従来，自然音声データの分析によって得られた韻律制御モデルや各音素の音響的特徴パラメータを用いて，出力波形そのものの音響的特徴を計算し，合成音声を新たにコンピュータで作成していたが，現在は単位選択手法により，音声波形データはそのまま合成時にも利用され，波形接続型による合成音声が作成される．これはコーパスベース音声合成と呼ばれている．

コーパスベース音声合成と従来法音声合成との大きく異なる点は，パラメータの使い方，つまりその情報利用にある．従来法では，学習結果により音声特徴パラメータを予測したが，現在はその同じパラメータ組合せの利用により音声データベースから音声単位を検索し，波形サンプルをそのまま抽出する．音声データは，検討材料としての利用から，合成のための単位音そのものへの利用へと拡大され，その結果，この合成手法の違いは合成音声の自然性となって現れる．

人の発話音声は複数レベルの情報を同時に表現し，音韻情報以外に話者情報，発話意図・感情・態度などの情報も含む．従来法においては，言語的意味が持つ音韻情報を概ね制御でき，分かりやすい合成音声を実現できたが，話者特徴（非言語情報）及び意図的内容（パラ言語情報）などを表現することに限界があった．非言語情報の制御は上記の波形接続合成の特徴であり，話者音声データベースの交換により，ある特定話者の音響的情報(声質・発話速度・ピッチの高さや幅・性別・年齢・感情状態等の情報）を表現できる．言語情報は当然どの合成方法でも伝えられる．

波形接続型音声合成では，元音声の自然性がそのまま残るため，非言語情報やパラ言語情報も表現可能となる．制御パラメータは言語情報のみならず，発話情報すべてを定義することが必要となるが，ボトムアップ的な統計学習結果による制御の代りに，トップダウン的に，発話特徴の制約条件を利用し，単位選択を行う．コーパスベース音声合成では，それぞれすべての特徴を制御するのではなく，有効な制約条件のみで，最適な音素単位列を決

目　次

Nick CAMPBELL (株)国際電気通信基礎技術研究所ネットワーク情報学研究所
E-mail nick@atr.jp
Nick CAMPBELL, Nonmember (Network Informatics Laboratories, ATR, Kyoto-fu, 619-0288 Japan).
電子情報通信学会誌 Vol.87 No.6 pp.497-500 2004年6月

定する．つまり，出力音声の特徴を作成するより，この制約条件のフィルタで微妙なニュアンスまで合成音声のコントロールが可能となる．

3. コーパスとは何か

元ラテン語の単語である「corpus」（コーパス）は，英語で，“a collection of naturally-occurring spoken or written material in machine-readable form”[1]，つまり，「記録された自然に現れる発話やテキストの集積」となる．専門用語「コーパス」の基本条件は，“… that are in themselves more-or-less representative of a language”[2]，すなわち，「ある言語の見本となる物」であり，“… for the systematic study of authentic examples of language in use”[3]，すなわち「その言語のそれぞれの利用環境による機能的役割を例示するもの」である．

ここで“language-as-system”（言語構造）と“language-in-use”（言語機能）の違いについて述べる．前者は書き言葉だけを調べてモデル化できる情報，つまり「言語学」である．後者はそれぞれの利用環境（コンテキスト）に依存し，社会言語学や心理学も含まれる．つまり日常的言語使用についての分析である．また，前者は音韻論や文法論を含むが，後者はパラ言語情報や非言語情報も考慮する．コーパスは言語機能の例示である．音声合成の研究に使われている音声データは，「コーパス」と呼ばれているが，実際には「データベース」のみである[4]．これらのデータベースは音声のバリエーションを示すが，発話音声の日常的な使い方の観点から見れば，不十分なものとなる．それはなぜか．利用されている音声データベースのほとんどがプロやアナウンサーの声による朗読音声データであるためである．言語情報の対話音声における「環境による機能的役割」つまり，声の音サンプルは当然自然であるが，データは「naturally-occurring」でないため，パラ言語コミュニケーションにおけるサンプルにはならない．

4. 音声合成の利用について

音声データベースは合成専門家のニーズに基づいて設計されたものであるが，ここで一般ユーザのニーズを中心として考えたい．合成に利用される音声データベースは，音声工学の観点からとらえれば，音声の音響的特徴を例示するといえるが，いわゆる高度メディア社会の情報コミュニケーションを考えた場合，日常的に，人間の代りにしゃべる合成アプリケーションが必要となる．この観点から，現在利用されているデータベースはまだ不十分である．特に，聞き手にとって情報不足である．

音声合成技術は歴史的に見れば，まず音声生成を目的として，声帯・声道の特徴，音声スペクトルの特徴，母音や子音などの特徴を検討し，音声学的な研究であった．つまり，“sound as system”といえる．第二フェーズはアクセントやタイミング等，韻律特徴の検討の段階に入り，文の区切り・強調などを定義した．これにより言語的意味の制御ができ，音声合成は reading machine（朗読機）となった．しかし人間の音声コミュニケーションという段階では，朗読だけの文読み上げにとどまらず，日常的に音声を合成する talking machine（しゃべり機）が求められる．

4.1 人間の声の代りに

近年，音声合成の自然性を高めるために，パラ言語情報を考慮するために，多くの研究者が「感情音声」をキーワードとした．しかし感情は非言語情報と考えられる．感情表現はパラ言語情報として有効であるが，喜怒哀楽で感情を定義するのは適切だろうか．例えば，発話者が風邪をひいた・酔っぱらった・熱があるなどと同様に，喜怒哀楽は話者状況情報で，非言語情報である．この情報とパラ言語コミュニケーションとの関係は少ない．しかし，ある話者が意図的に風邪をひいたように，酔っぱらったようにしゃべれば，その情報はコミュニケーションに直接関係があり，発話内容の意味判断に重要な役割を持つ．

音声合成は笑う必要も泣く必要もないといわれているが，音声合成の利用者，例えば，自分の声を自分でコントロールできない人も笑いたい，泣きたいなど，いろいろな感情表現を通して発話意図を示したい．音声障害者のみならずとも，多くの人が合成音声を使って遊びたい，商業にも使いたいと考えるであろう．実際の人の声の代りに合成を利用する場面はかなり多く考えられる．音声翻訳・ロボット・ゲーム・情報提供・カスタマケアなどの利用が既に提案されているが，それらはすべてパラ言語情報が必要である．現在の音声合成は，Turing Test（人間との区別ができないほど）はまだ通れない．

4.2 対話音声とパラ言語情報

人間の声が持つ情報は，言語情報（文内容），非言語情報（話者自身）のほか，パラ言語情報（話者意図，態度，感情，発話行為，相手との人間関係等の情報）が含まれ，このパラ言語情報は自然対話発話の約半分以上の情報量である[5]．講義やニュースを書き起すと，その文字列からほとんどの意味が伝えられるが，日常会話の発話音声の書き起しからは，言い方による意味の違いが含まれていないため，話者意図など文字列だけで表現されない情報が欠落する．対話音声では非語彙的情報が多い．例えば，「えー」，「あー」，「あ」，「あっ」，「へー」などの単音で話者状況，意図，意見などが示される．語彙的

情報よりも単音が持つ情報は韻律の変化が多く，相互的な発話行為の中で，声質のコントロールなどによって微妙なニュアンスの違いを表現できる．現時点の音声合成技術では，まだ，この段階の情報を伝えられない．

ここで，言葉と意味の関係に戻る．ある発話を言い換える，つまり別の文表現に置き換えたり，翻訳した場合，その文形式は変るが話者意図はそのまま残る．それに対して，言い方を換えれば，言葉はそのまま残るが，意味や意図が大きく変る．対話ではこのような「言葉の遊び」が人間の特徴の一つである．言葉の選択，言い方の選択，特に上述のような非語彙的短発声で意見や意図を示す．同じ「ア」でも，数種類の意味が言い方の違いによって表現できる．現在の音声合成にこのような自由度はまだない．ある一つの文字列を入力すると，一つの音しか生成されない．それは合成に必要な入力情報の規定ができていないこと，合成のテキスト処理モジュールは意図判断ができないこと，音声自身のバリエーションが足りないことによる．声の柔らかさ（裏声・地声など）のコントロールがなくて，音色における自由度のない技術である．

4.3 自然音声合成

最近よく使われるもう一つのキーワードは「spontaneous」（自発的音声）である．これは朗読音声に対するもので，spontaneous な音声の定義は「考えながら」や「発話内容を作成しながら」となる．しかし，人が発声するという上で，朗読であっても，自発的であるといえるため，ここで新たな定義を提案したい．それは「相互的」である．つまり，相手の反応を理解しながら発話様式を実時間で決定する発声の仕方．単なる情報提供の場合は一方向的コミュニケーションであり，相手に依存しない形で文内容のみをメディア変換して伝える．対話の場合，相手の判断を認識しながら自分の言い方や表現形式（声質も含む）を調整し，相互的コミュニケーションを行う．コンピュータが，相手の認識結果を判断するのは，まだ当分の間，不可能であるが，上記の「人間の代りに」のような利用においては，認識結果に基づいて次発話作成の入力パラメータを制御する必要がある[5]．

そのコントロールパラメータを決定するために大規模対話音声コーパスが不可欠となる．トップダウン的に，音声コーパスを設計・作成すれば，発話環境は自然性が制約される．ボトムアップ的に収集し，自然性を最大限に広げる．こうして作られたコーパスは人間同士の一般的な声や音声の使い方を含む．この考え方に従って作成されたものとして JST の ESP コーパス[6]がある．本コーパスは発話内容や発話様式を規定せず，長時間の自然対話を集めたもので，Labov による Observer's Paradox などをクリアした純対話音声である．声や表現形式のバリエーションが多く，spontaneous な音声の集積で，このコーパスからコーパスベース音声合成の開発が始まっている．

これまでの波形接続型音声合成に利用されている音声データベースのほとんどは，音韻バランスや韻律バランスをとるために，新聞や雑誌から選択された文書の朗読音声を用いていた．それに対して ESP コーパスは文字言語とかなり違う特徴を示し，音声言語のみのものであり，双方向的な相互コミュニケーションが多い．このコーパスは，語彙数が比較的少なく，単音の繰返しが多い．日常の対話であり，文構造が比較的単純でありながら，韻律や声質による意味やニュアンス表現はより細かくコントロールされている．この表現形式によって，意図・態度・感情ラベルの言い方ラベルを付与し[7]，ハイブリッドな波形接続合成手法を検証中である[8]．音素単位のみならず，フレーズ単位や独立発話単位をも利用した波形接続法となる．この方法では，まず，同一話者による朗読音声の音韻バランス音声データベースからターゲット音声を従来法 CHATR によって作成し，第二フェーズでは感情表現を含む大規模音声データベースから，この音響ターゲットに似ている音声候補ベクトルを選択し，候補発話単位を表現形式でフィルタリングし，最適な音声サンプルの接続で出力波形を作成する．この方法では，音響的制約条件による単位選択が行われ，音素ラベル情報が必要なくなるが，その代りに，表現形式フィルタが必要である．これは言い方ラベルから作成される（図1）．

音声合成は reading machine（朗読機）の場合，その入力は文字列だけとなり，すべての読み情報やアクセント情報・区切り情報などをテキストのみから予測する必要がある．一方，talking machine（しゃべり機）の場合には，テキスト入力とともに，意図情報や行為情報が含まれてないと，適切な発話様式の選択が不可能になる．しかし，人間のような相互的コミュニケーションでは，相手との人間関係情報，内容に対する興味度情報，発話者感情・意図・態度などの情報入力から表現形式特徴フィルタの制約条件を簡単に定義できる．この制約条件は言い方による意味やニュアンスの違いを制御する．

5. むすび

本稿では，まず，2種類のコーパス設計作成方法について比較した．トップダウンの場合，作成者の判断・設計により音声データベースとなるが，それは「コーパス」というより，「データベース」としかいえない．ボトムアップの場合，発話者自身がその音声のバリエーションを決め，実音声の例となる．このようなコーパスから，音声

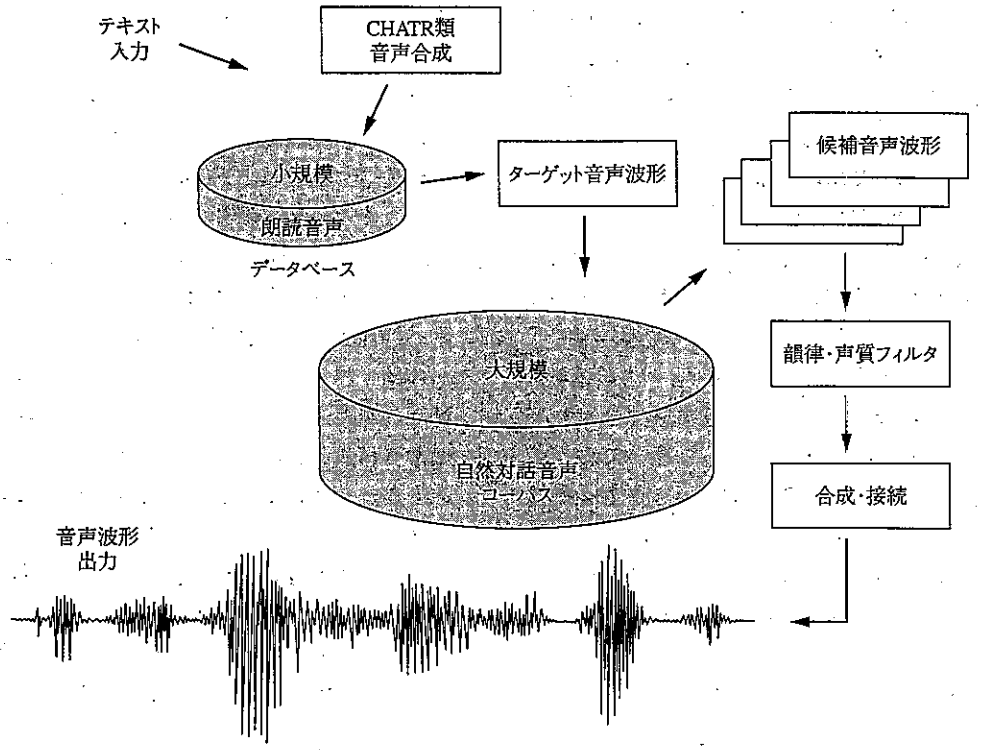

図1　ハイブリッド型音声合成の流れ　　テキスト入力からCHATR類合成器でターゲット音声をまず作成し，それを元に，大規模表現対話コーパスの複数候補より韻律・声質フィルタで最適な波形連結を出力する方法である．

合成を作成する場合，問題はかなり多いが，パラ言語的情報もその音声サンプルに含まれており，「人間の代りに」，次世代の音声コミュニケーションに利用する技術が可能となる．

コーパスベース波形接続型音声合成は，現在，音素単位を用いるが，今後，句単位や独立発話単位の表現形式がそのまま利用される．単位選択は現在音素環境と韻律環境に依存するが，将来の音声合成では，言い方環境・相手環境・発話意図・行為などの情報も含めて大規模自然発話コーパスから選択し，対話音声の合成が可能となる．

パラメータによる音声合成と比較すると，パラメータ自身はほぼ同じだが，そのパラメータの使い方が大きく異なる．従来法では予測結果から音響的特徴を近似したが，現在はパラメータによる制約条件で候補をフィルタリングする方法により，元音声の自然性や言い方のバリエーション，表現形式のコントロールが可能となる．

文　　献

（1） J. Sinclair, Corpus, Concordance, Collocation, OUP, 1991.

（2） The Oxford Companion to the English Language, McArthur & McArthur, ed., OUP, 1992.

（3） D. Crystal, A Dictionary of Linguistics & Phonetics, Blackwell (3rd edition), 1991.

（4） K. Hirose, personal communication, 15, Nov. 2002.

（5） W.N. Campbell, "What type of inputs will we need for expressive speech synthesis?," IEEE Speech Synthesis workshop, Santa Barbara, USA., Sept. 2002.

（6） JST/CREST Expressive Speech Processing project, introductory web pages at : www.isd.atr.co.jp/esp

（7） ニックキャンベル，"音声合成の観点から見た言語音声の特徴"，月刊言語，Oct. 2002.

（8） W.N. Campbell, "Towards synthesizing expressive speech; designing and collecting expressive speech data," in Proc Eurospeech 2003 (In Press).

Nick　CAMPBELL

現在，ATRネットワーク情報学研究所の主幹研究員・プロジェクトリーダで，科学技術振興事業団のCREST（ESP）表現豊かな音声情報処理プロジェクト研究代表者．英国のサセックス大，実験心理学博士．IBM（UK）科学研究センター，エジンバラ大CSTR，AT&T等の客員研究員．ATRは1990から，1998から奈良先端大の客員教授，1999から神戸大大学院連携講座の客員教授．研究課題は，大規模音声コーパスによる韻律特徴・音声合成・表現形式である．