相槌の音響的変化が対話の自然性に与える影響*

○森　大毅 (宇都宮大)

1　はじめに

人間味のある音声対話システムの実現を目指すとき，ユーザに向けてシステムが出力する音声応答には多くの検討すべき課題がある。対話音声の合成には，対話に特有のコンテキストになじむ自然性が要求されるため，明瞭性という言語的要素のほかに話し方のニュアンスというパラ言語的要素を考慮する必要がある。

本論文では，相槌音声の合成を論じる。機械との対話を自然で円滑なものとするために，適切な相槌の生成が重要であることは広く認識されている。中でも，相槌の発話タイミング制御については多くの研究がある。一方，相槌が伝達するパラ言語情報と相槌の音響的特徴との関係を分析した研究もある [1, 2] が，相槌音声の合成においてパラ言語情報を制御しようとする試みは多くない [3]。

本論文では，自然な対話において相槌が連続で打たれる場所に着目し，その音響的特徴を操作することで，対話全体の自然性がいかなる影響を受けるかを調査した知覚実験について述べる。著者らの仮説は以下の通りである。

仮説 1 全ての相槌の音響的特徴を同一にすると，対話全体の自然性が低下する。

仮説 2 相槌の音響的特徴を文脈に関係なく変動させると，対話全体の自然性が低下する。しかし，その低下の程度は全ての相槌を同一にした場合よりは小さい。

2　対話コーパスと相槌の音響的特徴

知覚実験で使用する音声刺激は，宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB)[4] に収録されている音声を利用して作成した。UUDB は，親しい友人同士の表情豊かな対話のコーパスであり，相槌の音響的特徴も多様である。

UUDB に収録されている対話では，相槌は比較的早いタイミングで打たれている。正常話者交替における先行発話の終端から当該発話の始端までの長さは平均 361 ms であるのに対し，相槌が打たれた発話断片の終端から相槌の始端までの長さは平均 55 ms である。また，相槌全体の 43%が相手話者の発話断片の終端より前に開始している。これは，相手発話の内容の深い理解を伴わず自動的に打たれている相槌が多いことを示唆している。

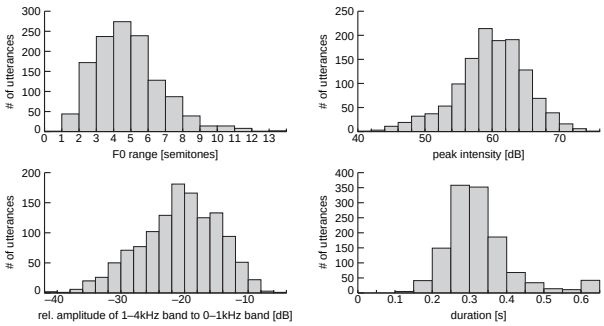

Fig. 1　発話「うん」の音響的特徴の分布

Fig. 1 に，UUDB 中の発話「うん」の音響的特徴 (F_0 レンジ，強度最大値，スペクトル傾斜，継続時間) の分布を示す。言語的内容が同じでも，音響的特徴が広い分布を持っていることがわかる。

3　実験方法

知覚実験で使用する音声刺激は，UUDB の音声を加工して作成した。まず，UUDB 中で，一方の話者 A の発話に対してもう一方の話者 B が連続で相槌を打っている部分を抽出した。以下，この部分のことを便宜的にターンと呼ぶことにする。ターンの認定は，以下の基準に従って行った。

- 話者 A の発話始端時刻から (発話終端時刻 + 500ms) の範囲に始端がある話者 B の発話「うん」が 3 個以上存在する

なお，UUDB では相手話者の発話末尾に対する応答「うん」は相槌とはみなされていないが，ここでは相槌と区別しないこととする。

このようにして，抽出された女性 6 ペアの 60 ターンの中から，極めて特異な相槌を含む・大きなノイズを含むなどの理由で実験に不適当だと判断されたものを除いた 51 ターンを呈示音声セットとする。呈示音声セットに含まれる発話「うん」の話者数は 10 名である。

次に，各ターンに含まれる「うん」の音声波形に対して 3 パターンの操作を加える (Fig. 2)。

加工なし 元の「うん」のまま

---

*On acoustic variations of aizuchi and the naturalness of dialogues.
by MORI, Hiroki (Utsunomiya University)

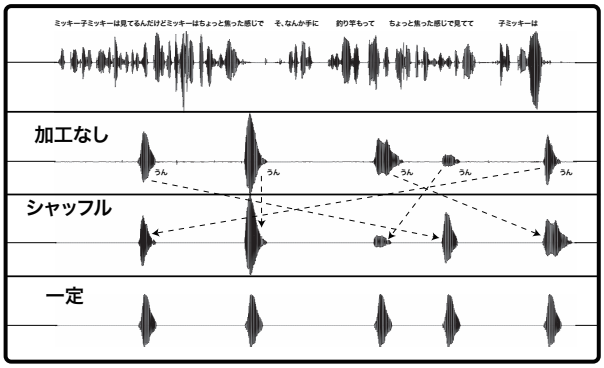

Fig. 2　相槌音声波形の操作

一定 各ターンの中の「うん」の波形を，全てどれか 1 種類の「うん」で置き換え

シャッフル 各ターンの中で,「うん」の波形をランダムに入れ換え

呈示音声セットに含まれる各ターンに対し上記の 3 パターンの操作を施した計 153 音声を刺激として実験に用いた。

被験者にはヘッドフォン (MDR-Z600, SONY) を装着させ，一方のチャンネルから相槌を打つ側の話者の音声を，もう一方のチャンネルから相槌が打たれる側の話者の音声を呈示した。153 音声をランダムに呈示し，各音声呈示後に対話の自然性を評価させた。

被験者には，相槌に着目して対話の自然性を判定するよう教示した。この際，1 つ 1 つの相槌を分析的に評価するのではなく，一連の対話を全体として評価するよう強調した。評語は (0:不自然, 1:どちらかというと不自然, 2:どちらかというと自然, 3:自然) の 4 種類であり，以降の分析では便宜的に 0–3 に数量化した値を用いる。被験者は音声科学に特段の知識を有していない大学生 10 名 (女性 5 名，男性 5 名) である。

4　実験結果および考察

各操作に対する全被験者の平均評定結果を Fig. 3 に示す。「加工なし」に対し「一定」では不自然と評価されていることがわかる。また,「シャッフル」の平均評定はその中間にあることがわかる。分散分析の結果は，操作パターンの主効果が有意であり ($p < .05$), Bonferroni 法を用いた多重比較の結果,「加工なし」と「一定」の間，および「加工なし」と「シャッフル」の間には有意差が認められたが,「一定」と「シャッフル」の間には有意差が認められなかった。

「加工なし」の場合に比べ「一定」の場合に自然性が低下したことは，仮説 1 を支持する結果である。被験者の一部は，文脈に関係なくいつも同じ相槌が聞かれることに対して不自然さを感じ取っており，このことが結果に反映したものと考えられる。

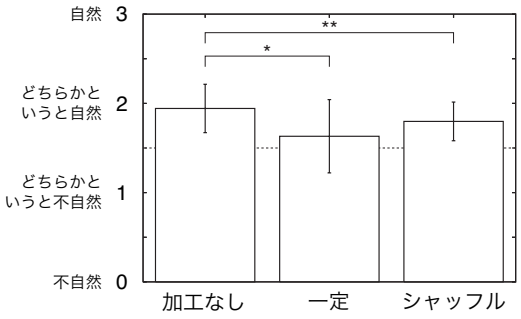

Fig. 3　全被験者の平均評定結果 (*:$p < .05$, **:$p < .01$)

また,「加工なし」の場合に比べ「シャッフル」の場合に自然性が低下したことは，仮説 2 を部分的に支持している。つまり，相槌の音は単に変化さえしていればよいわけではなく，相手話者の発話内容や音響的特徴などの文脈を考慮した制御をしない限り，自然性が低下する。

しかし,「一定」の場合に比べ「シャッフル」の場合に自然性が向上したとは統計的に認められず，仮説 2 が全体的に裏付けられたとは言えない。この原因として,「一定」に対する自然性評価の個人差が大きいことが挙げられる。平均評定結果の被験者別分布は,「加工なし」の最大 2.55，最小 1.47 に対し「シャッフル」が最大 2.39，最小 1.35 と大差ないのに比べ,「一定」は最大 2.35，最小 0.78 とばらつきが大きい。「一定」と「シャッフル」の間の自然性の違いの有無を精査するためには，被験者数を増やした上で類型化するなどの方策が必要と考えられる。

5　おわりに

音声対話システムが出力する相槌の音声合成におけるパラ言語情報の制御を目的として，自然な対話において相槌が連続で打たれる場所に着目し，その相槌の音響的特徴を操作することで対話全体の自然性が受ける影響を，知覚実験により調査した。同一の相槌が続く場合，および相槌をランダムにシャッフルした場合の両方で，自然性が低下することが確認された。

謝辞　研究協力者の斎藤佑樹氏ならびに長塚健二氏に深く感謝する。

参考文献

[1] 梅野他，音講論 (春), 313–314, 2003.
[2] 榎本, 石本, 情報処理学会研究報告, 2009-SLP-77, 1–6, 2009.
[3] 森, 音講論 (秋), 253–254, 2009.
[4] Mori et al., Speech Communication (in press).