Vol.2016-MUS-111 No.50
2016/5/22

音声における「かわいらしさ」の知覚と聴取時間の関係性の検討

大野 涼平[1,a)]　森勢 将雅[2,b)]　北原 鉄朗[1,c)]

概要：テキストを読み上げる音声合成技術において，利用者にとって魅力的に感じる音声，特に「かわいらしい」と感じる音声を自由に生成することはエンターテインメントの新たな創出には重要である．しかし，「かわいらしさ」の知覚と音響特徴量の関係は未だ十分に明らかにされていないだけでなく，人間が魅力的か否かを判断するために必要な発話の持続時間についても明らかにはされていない．そこで本稿では，「かわいらしさ」の知覚は音声の持続時間が短いと困難であるという仮説を立て，「かわいらしさ」の知覚にはどの程度の持続時間が必要か聴取実験によって調べた．実験の結果から，持続時間が短いほど判断にばらつきが生じることが明らかとなった．また，評価結果から，概ね安定してかわいらしさを評価するための持続時間についても議論する．

キーワード：知覚，音声，かわいらしさ，持続時間

1. はじめに

音声合成の研究は，当初は自然な音声を作ることに主眼が置かれてきたが，その後，感情音声のように表現豊かな音声の生成が取り組まれるようになり [1][2][3]，最近では，個々のユーザにとって心地よい，その人の好みに合った音声の生成も注目を浴びつつある．雑談のようなタスク指向型でない音声対話などでは，こういった主観的な印象が対話の継続性などに影響すると考えられるからである．その他，動画コンテンツの制作などにおいてもこうした技術は有用性が高いと考えられる．

個々のユーザにとって印象のよい音声を生成するには，そのような音声に対する印象が音響特徴量とどのような関係を持つかといった分析が不可欠である [5][6][7]．例えば，渋谷ら [5] は，親近感が高い発話は話者が主として使用する基本周波数（F0）の値とそれよりもやや高い値の間を滑らかに変化させているということを述べている．菅原ら [6] は，聴取者によって対象の音声が「いい声」かどうかの評価は異なるが，「いい声」と判断した音声には同じような印象を抱くということを報告している．また，音響的特徴量との関係性では，F0 平均値だけでなく話速度やスペクトルのピーク数にも関連性があるという可能性を示唆した．西田ら [7] は，印象表現語と韻律パラメータの相関関係を正準相関分析により学習し，発話印象の推定を行った．

個々のユーザの好みに合った音声を生成する 1 つの有力な方策が，そのユーザにとって魅力的な音声を生成することである．女性声に関して，近年，「萌え声」や「メイド声」の分析や社会科学的考察が増えつつある [8][9][10]．「萌え声」とは，特定の対象物に強い愛着を持つ「萌え」という俗語と「声」を組み合わせてできた言葉で，「かわいらしい声」に近い意味で用いられている．高野ら [8] は「萌え声」とは話者の声質に起因するもので，時間方向の F0 平均値や標準偏差，話速を操作することによって萌える声における萌え度の制御は可能ということを示した．「メイド声」はメイド（メイド喫茶で働く女声）が出す独特な声であり，川原 [9] によると，地声よりも仕事中の演技声の方が F0 平均値が高く，F0 の上昇時のその上昇量が大きく，話速は少し早い傾向にあるということを示している．このように，「かわいらしさ」と音響特徴量の関係は分析されているものの，未だ十分にされていない．

本研究では，音声に対するかわいらしさの知覚と聴取時間の関係，すなわち，何秒聴けばかわいらしさを知覚できるかについて調査する．もし，F0 の変化を知覚できないほどの短時間聴いただけでもかわいらしさを知覚できるとすれば，かわいらしさは音声のスペクトルに起因すると考

[1] 日本大学
Nihon University
[2] 山梨大学
University of Yamanashi
[a)] ryouhei@kthrlab.jp
[b)] mmorise@yamanashi.ac.jp
[c)] kitahara@chs.nihon-u.ac.jp

表 1 各声優における音声データの内訳

持続時間	切り出し位置	計	全声優での合計
75 [ms]	ランダムに 3 種	3	36
150 [ms]	ランダムに 3 種	3	36
300 [ms]	ランダムに 3 種	3	36
600 [ms]	ランダムに 3 種	3	36
フル	1 種（切り出しなし）	1	12
合計		15	156

えられる．それに対し，一定の長さを聴かないとかわいらしさを知覚できないのであれば，スペクトルや F0 の時間的な変化も，かわいらしさの知覚に重要であることが示唆される．

2. 実験条件

2.1 音源データ

本研究では「お兄ちゃん CD」[11] に収録されている音声を使用する．お兄ちゃん CD とは，12 名の各声優が 100 種類のシチュエーションで「お兄ちゃん」と発話した音声 (全 1200 音声) が収録された CD である．今回の実験では，聴取実験 [8] によって「萌え度」が高いという評価を得た発話から 12 人の全声優の音声を使用する．各声優の発話音声から 75 [ms], 150 [ms], 300 [ms], 600 [ms] 間を切り出す．このとき，切り出す位置はランダムに 3 箇所用意した (表 1)．

2.2 実験手順

被験者 14 名（大学生の男性・女性各 7 名ずつ）に次の通り実験を行った．

(1) 75 [ms] の音声 36 種類を 3 個ずつ含めた計 108 個の音声をランダム順に 1 回ずつ聴かせ，5 秒間でその音声に対する評価を 5 段階（1:「かわいらしさを感じない」, 5:「かわいらしさ」を感じる）で評価させる．108 個の中に 36 種類の音声を 3 個ずつ含めたのは，1 人の被験者による同一音声への評価の最大値と最小値の差がどの程度なのかを見るためである．

(2) 150 [ms], 300 [ms], 600 [ms] のときも (1) と同様に行う．

(3) 切り出しをしないオリジナルの音声（フル音声）12 種類を 3 個ずつ計 36 個の音声をランダムの順で 1 回ずつ聴かせ，(1)(2) と同様に評価させる．

3. 分析方法

本研究では，かわいらしさを知覚するには一定の聴取時間が必要で，聴取時間が短いほどかわいらしさを知覚できなくなるとの仮説を立てる．かわいらしさを知覚できない状態は，「3」を回答する他，同一の音声に対して回答が「1」から「5」までばらつく場合も考えられる．

3.1 分析 1

本実験では，同一条件の音声を各被験者が 3 回聴取して回答させることとした．聴取時間が一定時間を下回ると判断が困難になる仮説が支持される場合，3 回聴取した結果のばらつきは，聴取時間に反比例することとなる．この仮説を検証するため，分析 1 での分析とする特徴量として，3 回の評価値の最大値と最小値の差を用いることとする．以下では，この差をレンジと定義し用いることとする．音声の持続時間とレンジとの関係性を評価することで，同一音声に対して安定して判断できるために必要な時間を推定することが可能になる．分析 1 では，このレンジを特徴量とした解析を中心に行う．

3.2 分析 2

あるフル音声（切り出しなし）に対する結果が「4」あるいは「5」だった場合，その音声の持続時間を短くした切り出し音声に対しても同じく「4」または「5」と答えていたとは限らない．聴取時間が一定時間を下回ると判断が困難になる仮説が支持される場合，聴取した結果に「1」〜「3」が含まれてる割合は，聴取時間に反比例することとなる．この仮説を検証するため，分析 2 での分析とする特徴量として，3 回の評価値の中央値を用いることとする．以下ではこの中央値を評価値と定義し用いることとする．音声の持続時間と評価値との関係性を評価することで，持続時間が異なる同一音声に対して安定して判断できるために必要な時間を推定することが可能になる．分析 2 では，この評価値を特徴量とした解析を中心に行う．

4. 分析結果

4.1 分析 1

4.1.1 全データに対するレンジの分布

各被験者が 75 [ms] の音声（36 種類）に行った 3 回の評価におけるレンジを算出し，それを 75 [ms] における分布とする．各持続時間について同様の処理を行い，持続時間毎の分布を求める．各分布におけるレンジの累積確率が図 1 上段である．

コルモゴロフ・スミルノフ検定（KS 検定）の結果から 5 つの分布に正規性が見られなかったため ($p < 10^{-30}$)，クラスカル・ウォリス検定（KW 検定）を行った．その結果, 5 つの分布のうち少なくとも 1 組の 2 分布間においては中央値に有意な差があるということがわかった ($p = 6.045 \times 10^{-8}$)．5 分布間でマン・ホイットニーの U 検定（U 検定）による多重比較を行った結果を図 1 下段に示す．75 [ms] の分布は 150 [ms] 以外の分布より有意に中央値が高いということがわかる．このことから，75 [ms] の音声に対するレンジは，少なくとも 300 [ms] 以上の音声より大きいことがわかる．つまり時間が短くなるにつれて同一音声に対する評価のばらつきが大きくなるということを示す．

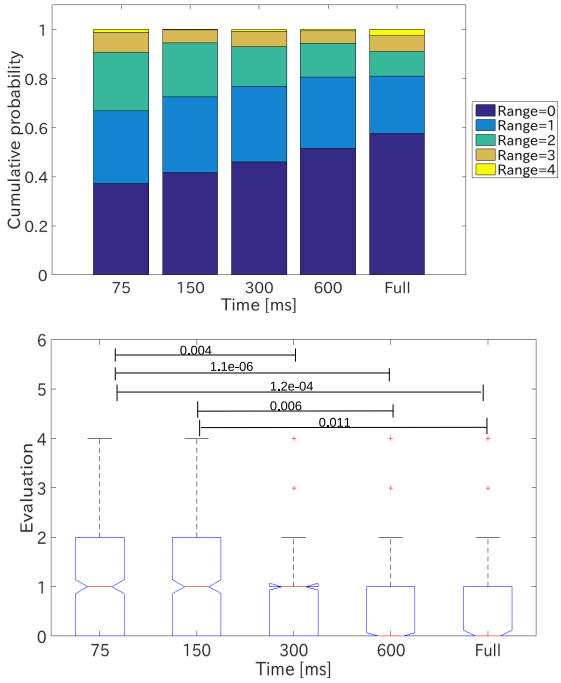

図 1 レンジの累積確率（上段），分布の箱ひげ図（下段）
下段の横線上の数値は p 値を表す

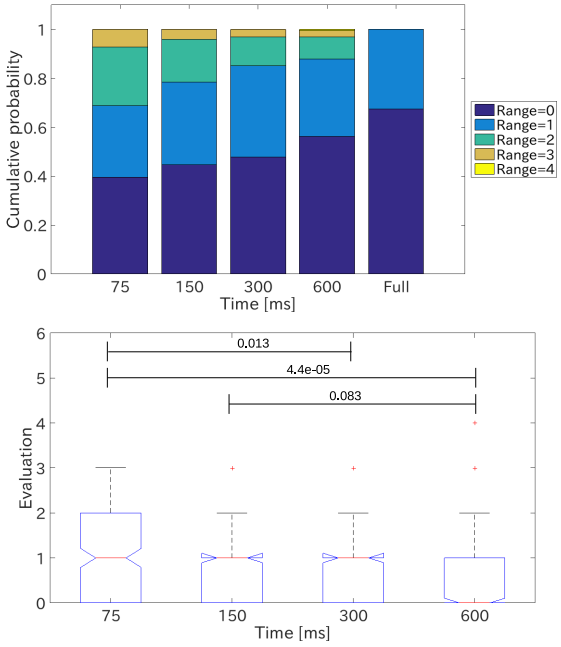

図 2 「かわいらしい」と答えた音声におけるレンジの累積確率（上段），分布の箱ひげ図（下段）

4.1.2 「かわいらしい」と答えた音声におけるレンジの分布

フル音声に対する評価で 3 回とも「4」以上と回答した音声に対する 3 回の評価におけるレンジを分布とする（累積確率は図 2 の上段）．

{ また，以降の検定では，フル音声に対する評価の分布を除いた 4 分布における分析を行う．これはフル音声における分布は評価が 4 以上のものを取り出しており，3 以下のものを含まない特殊な分布となるからである．}

KS 検定ではどの分布でも $p < 10^{-49}$ で，正規性が見られなかった．KW 検定では $p = 5.1 \times 10^{-5}$ であった．また，多重比較の結果を図 2 の下段に示す．時間が短くなるにつれ中央値が高くなっており，これは時間が短くなるにつれて同一音声に対する評価のばらつきが大きくなるということを意味する．

4.1.3 「かわいらしくない」と答えた音声におけるレンジの分布

フル音声に対する評価で 3 回とも「2」以下と回答した音声に対する 3 回の評価におけるレンジを分布とする（累積確率は図 3 の上段）．

KS 検定ではどの分布でも $p < 10^{-35}$ で，正規性が見られなかった．KW 検定では $p = 0.190$ で，分布間の中央値に有意な差があると言えなかった（図 3 の下段）．このことから，フルのとき「かわいらしさ」を感じない音声に関しては時間長と同一音声に対する評価のばらつきには関係性があるとは言えない．

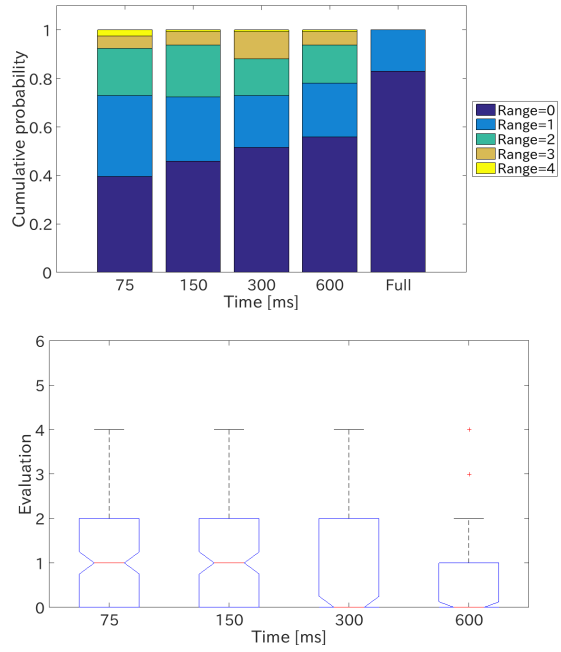

図 3 「かわいらしくない」と答えた音声におけるレンジの累積確率（上段），分布の箱ひげ図（下段）

4.1.4 分析 1 のまとめと考察

持続時間が短くなるにつれ，分布の中央値が高くなるということが言えた．従って，持続時間が短ければ同一音声に対する 3 回の評価はばらつきが大きくなると言えた．一方でフル音声に対する評価で 3 回とも「4」以上と答えた音声では，持続時間が短くなるにつれてレンジ $= 2, 3$ の確率が増えていき，3 回とも「2」以下と答えた音声では，持続時間に関係なく一定であった．また，分布の中央値も 3 回

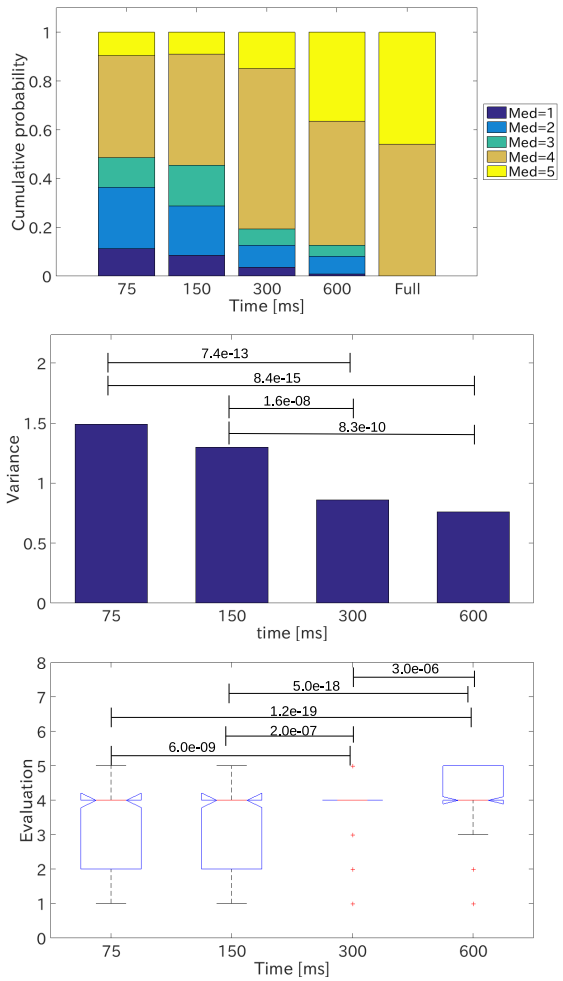

図 4 「かわいらしい」と答えた音声における中央値の累積確率（上段），分散（中段），分布の箱ひげ図（下段）

とも「2」以下と答えた音声では持続時間の違いによって有意な差はなかった．以上のことから，フル音声で「かわいらしい」と感じる音声の評価は持続時間が短くなるにつれて評価の安定性にばらつきが大きくなるが，「かわいらしい」と感じない音声の評価は安定性は変わらないということがわかった．また，150 [ms] から 300 [ms]，300 [ms] から 600 [ms] の間で分布の中央値に差が見られた．

4.2 分析 2

4.2.1 「かわいらしい」と答えた音声における中央値の分布

フル音声の聴取結果で 3 回とも「4」以上と回答した音声に対する 3 回の評価における中央値を分布とする（累積確率は図 4 の上段）．

KS 検定ではどの分布でも $p < 10^{-100}$ で，正規性が見られなかった．4 分布間における等分散性を検定するためにルビーン検定を行った結果，$p = 5.1 \times 10^{-21}$ で少なくとも 1 対の分布間で分散が有意に異なるとわかった．その後，多重比較を行った結果，図 4 の中段のようになった．また，KW 検定の結果は $p = 2.8 \times 10^{-27}$ となり，多重比較の結果は図 4 の下段のようになった．分散に関しては，時間が短くなるにつれて分散が大きくなっており，評価のばらつきが大きくなっているとわかる．分布に関しては，75 [ms] と 150 [ms] の分布間以外全ての組で有意な差があり，時間が短いと中央値が「4」，「5」より小さくなり，「4」や「5」以外の評価が多くなると言える．

しかし，それでも短い時間においても 4 以上の評価が 5 割程度であった（図 4 上段）．このことから短い音声の中で何かしらの特徴量を手がかりに聴取者が話者の音声の全貌を予想して回答している可能性があると言える．一方で，75 [ms]，150 [ms] では 4 以上の評価が 5 割だが，時間が短いと「かわいらしい」と感じる音声も「かわいらしくない」と判断してしまうということがわかる．

4.2.2 「かわいらしくない」と答えた音声における中央値の分布

フル音声に対する評価で 3 回とも「2」以下と回答した音声に対する 3 回の評価における中央値を分布とする（累積確率は図 5 の上段）．KS 検定ではどの分布でも $p < 10^{-100}$ を下回り，正規性が見られなかった．またルビーン検定の結果，$p = 1.8 \times 10^{-5}$ となり，多重比較の結果は図 5 の中段のようになった．KW 検定の結果は $p = 0.0478$ であったが，多重比較の結果はどの対でも有意な差はみれなかった（図 5 の下段）．

分散に関して，持続時間と分散の大きさは単調増加（減少）という関係ではなかった，分布の累積確率を見ると，中央値 $= 1, 2$ の確率は 75 [ms]，150 [ms]，300 [ms] とで変わらず，中央値 $= 4$ の確率が 300 [ms] のときに増えた．これが分散を大きくしたことに影響している．

平均順位に関しては，短い時間でも過半数の回答が「2」や「1」であった（図 5 上段）ことからも有意差がでなかったと考えられる．

また，3.2.1 節では反対の評価の確率が持続時間が短いと 0.5 に近づくという傾向だったにも関わらず，今回は持続時間が短い分布においても「2」以下の確率が 0.7 程度もあり，150 [ms] から 600 [ms] における「2」以下の確率は変化していない．これは，「かわいらしさ」を感じなかった場合に「3」ではなく 2 以下の評価が多く集まったのだと考えられる．

4.2.3 分析 2 のまとめ

3.2.1 節では「2」以下の累積確率が，時間が短くなるにつれて大きくなるのに対し，3.2.2 節では「4」以上の累積確率が大きくなるということはなかった．従って，持続時間が短くなるにつれフル音声における 3 回の評価が 3 回とも「4」以上と回答した場合でも，「3」以下と回答する割合が増した．一方で，持続時間が短いにも関わらず 3 回とも「4」以上あるいは「2」以下と回答している音声も多かった．つまり，持続時間が短くても「かわいらしさ」の判断自体は可能ではあるが，持続時間が増加するにつれて，フ

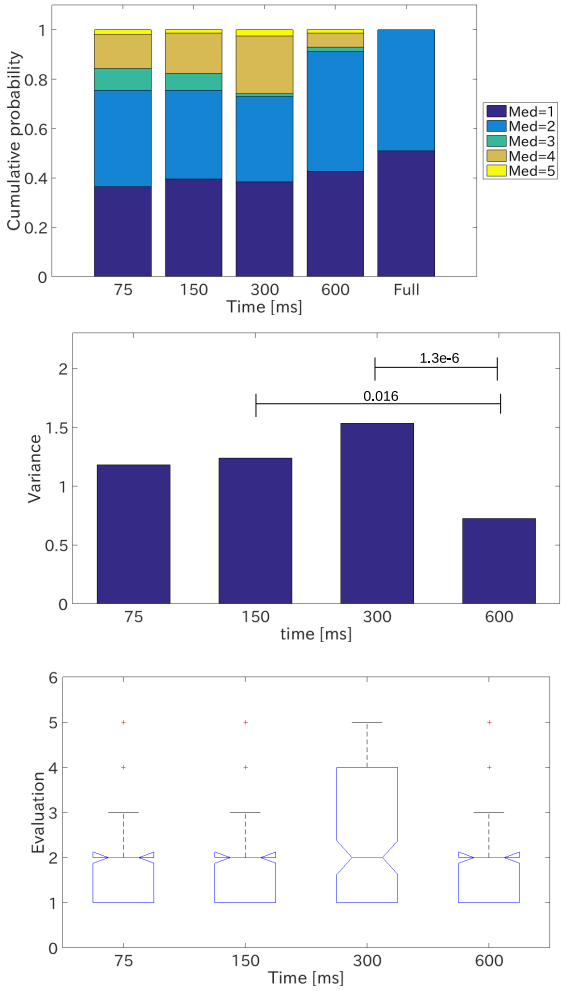

図 5 「かわいらしくない」と答えた音声における中央値の累積確率（上段），分散（中段），分布の箱ひげ図（下段）

ル音声では「かわいらしさ」を感じる音声に関しては 3 回とも「4」以上と判断を変更すると言えた．また，150 [ms] から 300 [ms]，300 [ms] から 600 [ms] の間に分布の分散や中央値に有意な差が見られた．

5. おわりに

今回は聴取実験によって聴取者の「かわいらしさ」に対する評価と聴取時間の関係性について調べた．聴取時間が長くなれば各被験者毎の評価の安定性が高くなるということがわかった．また，フル聴取時に「かわいらしい」と評価した音声は，短くても過半数が「かわいらしい」という評価であり，逆に「かわいらしくない」と評価した音声も同様で，短くても過半数が「かわいらしくない」と評価している．また，150 [ms] から 600 [ms] の間で評価値の安定性が変化していることはわかったが，間の 300 [ms] が境とは言えない．これらから言えることは，短い音声でも何かしらの特徴量を手がかりにその音声の評価をつけることは可能かもしれないということである．今後は短い音声に評価を付ける際，どのような特徴量を手がかりにしているかを検討していきたい．

謝辞 本研究の一部は，JSPS 科研費 16H05899 の助成を受けたものである．

参考文献

[1] 森山 剛，森 真也，小沢 慎治：“韻律の部分空間を用いた感情音声合成”，情報処理学会論文誌，Vol.50，No.3，pp.1181-1191，2009.
[2] 佐藤 賢太郎，広瀬 啓吉，峯松 信明: “生成過程モデルに基づくコーパスベース感情音声合成とその評価”，情報処理学会研究報告音声言語情報処理，pp.51-56，2004.
[3] 河津 宏美，長島 大介，大野 澄雄: “生成過程モデルに基づく感情表現における F0 パターン制御規則の導出と合成音声による評価”，電子情報通信学会論文誌，Vol.J89-D，No.8，pp.1811-1819，2006.
[4] 神山 歩相名，篠崎 隆宏，岩野 公司，古井 貞熙: “自然生と個人性に優れた音声合成のための音素継続時間長適応法”，日本音響学会春季講演論文集, No.2-7-1, pp.329-330, 2010.
[5] 渋谷 貴紀，川端 豪:“音声対話システムのための親近感特徴量の探索”，電子情報通信学会技術研究報告，Vol.106，No.122，pp.25-30，2006.
[6] 菅原 衣織，伊藤 貴之: “女性の声を例にした音響的特徴量と印象評価の関係性の調査”，エンターテインメントコンピューティング論文集，pp.67-72，2014.
[7] 西田 昌史，小川 純平，堀内 靖雄，市川 熹: “対話音声を対象とした韻律情報による発話印象のモデル化”，電子情報通信学会技術研究報告，Vol.105，No.493，pp.79-84，2005.
[8] 高野 佐代子，竹澤 勇希，竹内 純基，山田 真司: “「萌え声」心理的評価，音響分析および STRAIGHT を用いた合成音声評価”，日本音響学会春季講演論文集，No.2-Q5-22，2014.
[9] Kawahara,Shigeto: “The phonetics of Japanese maid voice I: A preliminary study”，pp.19-28，2013. Phonological Studies,
[10] 松原 実香，サトウ・タツヤ: “対象，評価，情動の観点から検討する「萌え」”，立命館人間科学研究，Vol.26 pp.21-34，2013.
[11] Cffon（レーベル）:“お兄ちゃん CD”，2006.