社団法人　日本音響学会

<http://www.asj.gr.jp/qanda/answer/50.html>

CiNii　論文引用

　子供の音声は，声道長や基本周波数が成人音声と異なるだけでなく，発声スタイルが自発的で年齢や個人による声の変動も大きいため，通常の成人用音韻モデルでは認識精度が著しく劣化する．また，子供が正確に文章を読み上げることは多大な労力が必要であり，大規模な整った音声データベースの作成が難しい．本研究では，子供の自発的な発話の高精度な認識を目指して，音声情報案内システムによる子供の実音声の大規模収集，年齢層別子供用音韻モデルの構築と評価，および教師なし話者適応の検討を行った．大語彙連続音声認識実験より，実環境で収集した子供音声を用いることで，単語認識精度が71.1%と既存の読み上げ音声モデルに比べて絶対値で23.9%の改善が得られた．また，年齢層別の傾向では，特に幼児の音声において年齢層依存モデルによる大幅な精度改善が見られた．次に，自動収集した話者ラベルなしの大量データに対する，自動話者クラスタリングを用いた十分統計量に基づく教師なし話者適応を提案した．提案法により59 966個の発話データをクラスタリングし，近傍話者クラスタを用いて音韻モデルを適応することで，クラスタ数200の条件において，年齢層依存モデルに対してさらに幼児で2.2%，低学年子供で1.7%，高学年子供で0.5%の認識性能の改善が得られた

（1回目）月読アイの認識テストを終えて

<結果>

　・動物英単語の最も認識率の良かったパラメータは”速さ½”の43％で、英単語は”cow”の76％であった。

　・飲食物英単語は全体的に認識が悪く、最も良いパラメータは”音程変化”の20％で、英単語は”hamburger”の48％であった。

　・動物英単語と飲食物英単語の認識率に大きな差が生じた。環境、使用音源はすべて同じにした。

<原因>

　・iPodtouchに音声を入れその音声をマイク入力でjuliusに認識させたから？(余計なノイズが入っていた？)

　・単語辞書の中に似た音をもつ単語がはいっており、誤認識がおこったのではないか

　・マイクデバイスそもそもの集音効果が悪かったのではないか

　・月読アイの音声波形または周波数がjuliusが認識しにくいものだったのではないか

<改善>

　・日本語英語の入力ができる証明→つまり、自分の声である程度の認識率を確保すればjuliusの認識は正しい。そして、認識率の悪さは小学生の音声周波数に合わせた音響モデルを作成することで改善できる。

　・音が被らないように単語辞書を調整する。

　・マイクデバイスを良いものに変える。

＜単語別の誤認識の例＞

・salmon

hamburger,cow

・fish

　語尾を伸ばしているので、単語辞書の修正で改善できるかも

・

(2回目)

<認識率５０％未満の英単語>

　・2回目行った理由・・・1回目の認識テストの際、誤認識が目立ったがどの英単語と誤認識していたのかを明　　　　　　　　　　　　　　　　確にしておらず、分析が行えないと考えたため。

　・動物英単語で最も良かったパラメータは最適化と速さ1/2の”43％”だった。最も良かった英単語は”cow”の96　　　％であった。

　・飲食物英単語で最も良かったパラメータは音程変化の”26％”だった。最も良かった英単語は"salmon"の60％　　であった。

　・1回目と比べてみると、動物英単語のパラメータのパーセンテージに変化はないが、英単語の認識率に大き　　な向上が見られた。そして1回目結果が芳しくなかった飲食物英単語は、同じく音程変化がパラメータ中最も　　よく、6%ほど改善がみられた。

動物英単語の読みを読みと一音目の子音と、発音をローマ字にした時の母音を抜き出した。(ん、小文字、伸ばしは考えない）結果をいかに示す。

単語名 認識率 子音 母音 誤認識回数と％

・cat（ｷｬｯﾄ） 28 C a,o 17回　cow（29％）、wolf（41％）、water（5％）、lion（17％）

・cow（ｶｳ） 96 C a,u 1回

・giraffe(ｼﾞｭﾗﾌ） 60 G u,a, 10回　wolf（20％）、horse（50％）、lion（20％）

・bear（ﾍﾞｱ） 8 B e,a 23回　wolf（52％）、horse（21％）、hamburger（4％）、　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　 　　　　認識せず（21％）

・monkey（ﾏﾝｷｰ）16 M a,i 22回　horse（90％）、salmon（13％）、cow（4％）、　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　 giraffe（4％）、candy（4％）

・wolf（ｳﾙﾌ） 24 W u,u,u 19回　horse（63％）、cow（10％）、tiger（26％）

・tiger（ﾀｲｶﾞｰ） 0 T a,i,a 25回　wolf（20％）、horse（88％）、salmon（4％）

・lion（ﾗｲｵﾝ） 16L　　　a,i,o　　 　　　　21回　wolf（95％）、horse（14％）、hambuger（4％）

　この結果を見ると、認識率に影響しているのは母音の”u”ということがわかる。また、”a,u”の組み合わせはこのシステムにとって非常に認識しやすい音になっているようだ。反対にした”u,a”でも認識率60％を確保している。

"a,i”や"e,a”の組み合わせは、人間では聞きやすくてもシステムにとっては認識しにくい音になっていた。

子音については確定的ではないが、”C”から始まる英単語は認識しやすいということが読み取れた。反対に”T"や”B"は認識率が特別悪かった。

　次に誤認識について。総合的にみると"wolf"と"horse”がパーセンテージが高いものとなった。”horse”に関しては今回の認識テストにおいて必要ないので単語辞書から削除しておくべきである。必要である単語以外を登録すると思わぬところで障害があることがわかった。

飲食物英単語の読みを読みと一音目の子音と、発音をローマ字にした時の母音を抜き出した。(ん、小文字、伸ばしは考えない）結果をいかに示す。

単語名 認識率 子音 母音 誤認識回数と％

・hamburger(ﾊﾝﾊﾞｶﾞｰ) 52 H a,a,a 12回　candy（16％）、horse（16％）、wolf(25％)、　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　giraffe(8％)、cow(16％)、認識せず(41％)、　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　monkey(8％)

・water(ｳｫｰﾀｰ) 20 W o,a 20回　wolf(40％)、salmon(15％)、cat(5％)、　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　cow(30％)、giraffe(25％)、candy(5％)

・salmon(ｻｰﾓﾝ) 60 S a,o 10回　認識せず(100％)

・meat(ﾐｰﾄ) 4 M e,o 24回　horse(91％)、salmon(8％)

・candy(ｷｬﾝﾃﾞｨ) 4 C a,i 24回　horse(95％)、lion(4％)

・fish(ﾌｨｯｼｭ) 4 F i,u 24回　salmon(100％)

・yogurt(ﾖｰｶﾞｰﾄ) 0 Y o,a,o 25回　lion(8％)、cat(36％)、bear(8％)、cow(28％)、　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　wolf(16％)、giraffe(20％)、candy(12％)

　動物英単語で推定した結果通りとまではいかないが、"a"が母音の最初に含まれていると少なからず認識率があることがわかった。面白いことに母音の組み合わせによって認識率に大きな差がでていることがわかった。例えば、salmonの母音"a,o”とcandyの母音"a,i"を比べてみる。どちらも最初の母音は"a"であるにも関わらず、そのあとに続く母音で認識率が56％も違っている。これはあくまで推測だが、"a"、"u"はjuliusシステムの認識においてプラスになっていて、残る"i"、"e"、"o"はマイナスの効果になっていると考える。つまり母音の最初の音が"a"且つ、次に続く音がプラスの効果をもたらす母音”a”、”u”が入っていれば高い認識率を確保することができる、ということになるはずである。その証拠に、cow（母音”a,u”）は96％を出している。子音についてみてみると、動物英単語で最も認識率が高かった英単語の子音がCであったが、飲食物英単語の子音Cをもつ英単語の認識率をみると、4％と非常に低く子音は認識において全く関係ないといってもよいだろう。

　次に誤認識について。全体的に見ると、動物英単語の分析と大差はなかったが、”－”（伸ばし棒）の数が多い英単語の誤認識の種類がほかに比べると少し多いことがわかった。伸ばし棒が認識率に影響するのかもしれない。

各パラメータ（最適化、速度2倍、速度1/2、音量2倍、音程変化）ごとの集計結果を以下にしめす。

パラメータ　　 　　　平均認識率

　最適化　　　　　　 30.7％

　速度2倍　　　　　 21.3％

　速度1/2　　　　　 30.7％

　音量2倍　　　　　 24.0%

　音程変化　　　　　 24.0％

これを見ると、最適化、速度1/2が少し高いというだけで、ほぼ横並びなのがわかる。

結果から、認識においてはパラメータを変化させたところであまり大差がないということがわかった。

小学生に認識してもらう際、認識がうまくいかない場合はゆっくり話すなどの処置をとる程度でよいだろう。

認識率が悪い母音iとeを含む英単語　(7単語)

　・bear　-　今までの音声でOK

　・monkey　-　マとンの間の間隔が長かったので速度を速くし抑揚を少しつけた

　・tiger　-　タとイの間の間隔が長かったので速度を速くし抑揚を少しつけた

　・lion　-　速度を少し速くし、抑揚を少しつけた

　・meat　-　速度を少し速くし、抑揚をつけた

　・candy　-　抑揚をつけた

　・fish　-　速度を速くし、抑揚をつけた

認識率が悪い英単語(上記以外)目安50％以下　(4単語)

　・cat

　・wolf

　・water

　・yogurt

3回目の認識テスト結果

　2回目の認識テストで母音によって認識率に違いがみられた。そこで特に認識率の悪かった母音iとeを含む英単語をピックアップし、音声データから見直した。

　前回でパラメータによる大きな違いはないとわかったので、今回は元にする音声だけのテストを試行回数5回で行う。

　英単語 母音 5回の確率 誤認識の比率 前回比

　・bear e,a 　　 40% cat(100%) 40%↑

　・candy a,i 　　 0% cat(60%)、horse(40%) →

　・fish i,u 　　 60% horse(100%) 60%↑

　・lion a,i,a 　　 20% giraffe(100%) 20%↓

　・meat i,u 　 0% horse(100%) →

　・monkey a,i 　　 0% cow(80%)、lion(20%) 60%↓

　・tiger a,i,a 　　 0% giraffe(20%)、cow(20%)、認識せず(60%) →

　変化なしがcandy、meat、tiger。向上がみられたのがbear、fishの2単語。低下したのがlion、monkeyの2単語。向上にしても低下にしても大幅な数値の変化が見られた。低下したものと、前回と変わらず0%だった英単語は母音に必ず"i"が含まれており、"a,i”の組み合わせで始まるものがほとんどだった。向上したものの”fish”は"i"を含んでいるのにも関わらず、60%という高い認識率を得た。発音する際伸ばし棒が入っているcandy、meat、monkey、tigerは認識率が著しく低いことがわかる。

　2回目の認識テストの分析の際に推定していたとおり、母音iはマイナスの影響をもたらすことがわかった。また、認識しにくい単語の特徴として音を伸ばすものは避けた方が良いということもわかった。今回音声を見直しすることで改善がみられるかと踏んでいたが、修正していない(手抜きではなく、もともとの音声がネイティブに近かった)bearが向上していたので、またjuliusの認識基準がわからなくなった。使用しているプログラムを理解する必要があるかもしれない。

4回目の認識テスト(単語辞書での伸ばし音の使用禁止)

　3回目のテストの結果で伸ばし音が入っていることにより認識率が低下するのではないかという仮説が生まれた。そこで、その仮説が正しいのか検証するため、単語辞書から伸ばし音をあらわす記号「：」を使用しない書き方で修正を行った。例えばbearは通常単語辞書では「b　e　a：」と記述するが、「b　e　a　a」というように本来伸ばし音が入る音の前の母音（または子音+母音）の音を続けて記述することで疑似的に伸ばし音を再現した。

　4回目の認識テストの結果を以下に示す。試行回数は前回と同じ5回に設定した。

　英単語 母音 5回の確率 誤認識の比率 前回比

・bear e,a 0% giraffe(40%)、horse(60%) 40%↓

・candy a,i 0% cow(40%)、horse(60%) →

・meat i,u 20% horse(100%) 20%↑

・monkey a,i 0% cat(40%)、cow(60%) →

・tiger a,i,a 0% 反応なし(100%) →

結果から伸ばし音による認識率の低下は関係ないということがわかった。

ただし、ひとつだけ認識率が改善された音があった。meatであるが、この音だけ伸ばし音の位置が音と音の中間にあることがわかった。疑似的に伸ばし音をつくる場合は伸ばし音が英単語の中間位置にくる単語だけ適用させればよいのかもしれない。