人間 GAN：人間による知覚評価に基づく敵対的生成ネットワークと音声の自然性知覚における評価*

☆藤井 一貴（徳山高専，東大院・情報理工），齋藤 佑樹，高道 慎之介（東大院・情報理工）
馬場　雪乃（筑波大学），猿渡 洋（東大院・情報理工）

1　はじめに

機械学習における生成モデルは，メディア処理に関連する研究に強く貢献している．特に，深層ニューラルネットワーク (deep neural network; DNN) に基づく生成モデル（深層生成モデル）は，DNN の強力な非線形変換の恩恵を受けて複雑なデータ分布をモデル化できる．敵対的生成ネットワーク (generative adversarial network; GAN) [1] は深層生成モデルの代表例である．GAN の枠組みでは，生成モデルと別に識別モデルも利用する．生成モデルは識別モデルを詐称するように学習され，識別モデルは実在データと生成データを識別するように学習される．これらを繰り返すことで，最終的に生成モデルは実在データ分布に従うデータをランダムにサンプリングすることができる．この枠組みは多様な分野においてその適応成功例 [2] が報告されており，今後も多様なメディアに対する効果が期待される．

しかしながら，通常の GAN は実在するデータ（すなわち，学習データ）の分布しか表現することができない．一方で，人間の知覚は実在するメディアからの逸脱に対してある程度の許容範囲を持つ．例えば，人間の自然音声を加工して創り出された非実在の音声に対して，我々は人格・キャラクタ性を認めることができる．本研究では，当該メディアにおいて人間が許容できる範囲を知覚分布 (perception distribution) と呼び，この知覚分布を表現できる GAN を検討する．前述した音声の例のように，知覚分布が実在データ分布よりも広い領域をカバーすると仮定した場合，膨大な実在データを使用したとしても，識別モデルを詐称して学習を行う通常の GAN では，知覚分布を表現することは不可能である．

そこで，本研究では人間の知覚分布を表現することを目標として，人間の知覚評価を詐称して生成モデルを学習する GAN（人間 GAN）を提案する．人間 GAN では，人間を「生成データに対して事後確率（すなわち，生成データに対する許容度）の差分を出力する black-box システム」とみなし，人間による知覚評価を利用して生成モデルを学習する．**Fig. 1** に，通常の GAN と提案する人間 GAN の違いを示す．通常の GAN では，DNN で記述される識別モデルを詐称するのに対して，提案する人間 GAN では，クラウドソーシングサービスの人的リソースから得られる知覚評価を詐称する．本稿では，音声の自然性（すなわち，人間が当該音声を人間の音声としてどの程度許容できるか）における実験的評価を行い，(1) 実在データ分布と知覚分布が異なること，また，(2) 提案する人間 GAN は，通常の GAN では表現できない知覚分布を適切に表現できることを示す．

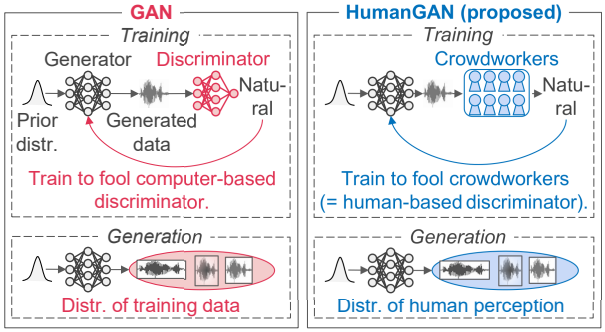

Fig. 1　通常の GAN と提案する人間 GAN の比較

2　通常の GAN

通常の GAN の目的は，生成モデルの表現する分布を学習に用いられた実在データの分布に一致させることである．そのため，通常の GAN は，データを生成する生成モデル $G(\cdot)$ と，実在データと生成データを識別する識別モデル $D(\cdot)$ を用いる．$G(\cdot)$ と $D(\cdot)$ は DNN で記述される．ここで，N 個の実在データを $\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \cdots, \boldsymbol{x}_N]$ とする．生成モデル $G(\cdot)$ は，既知の確率分布（例えば，一様分布）に従う N 個の乱数 $\boldsymbol{z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \cdots, \boldsymbol{z}_N]$ を，生成データ $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_n, \cdots, \hat{\boldsymbol{x}}_N]$ に写像する．識別モデル $D(\cdot)$ は，実在データ $\boldsymbol{x}_n$ もしくは生成データ $\hat{\boldsymbol{x}}_n$ を入力とし，当該入力が実在データである事後確率を出力する．学習時の目的関数 $V(\cdot)$ は次式となる．

$$V(G, D) = \sum_{n=1}^{N} \log D(\boldsymbol{x}_n) + \sum_{n=1}^{N} \log\left(1 - D(G(\boldsymbol{z}_n))\right) \tag{1}$$

2.1　生成モデルの学習

生成モデル $G(\cdot)$ は，Eq. (1) を最小化するように学習される．$G(\cdot)$ のモデルパラメータを θ_{G} とすると，θ_{G} は次式で推定される．

$$\theta_{\mathrm{G}} = \underset{\theta_{\mathrm{G}}}{\operatorname{argmin}} \sum_{n=1}^{N} \log\left(1 - D(G(\boldsymbol{z}_n))\right) \tag{2}$$

すなわち，$G(\cdot)$ は，$D(\cdot)$ が生成データを実在データとして識別するように学習される．一般的に，この計算過程は微分可能であるため，θ_{G} は逆誤差伝播法 (backpropagation) を用いて反復的に更新される．

---

*HumanGAN: generative adversarial network with human-based discriminator and its evaluation in naturalness modeling of speech, by Kazuki Fujii (National Institute of Technology, Tokuyama College, University of Tokyo), Yuki Saito, Shinnosuke Takamichi (University of Tokyo), Yukino Baba (University of Tsukuba), and Hiroshi Saruwatari (University of Tokyo).

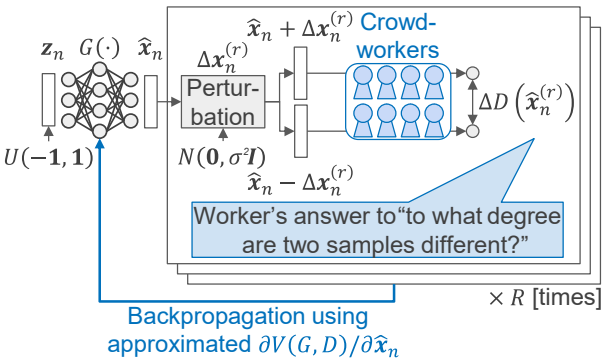

Fig. 2　人間 GAN における生成モデルの学習

2.2　識別モデルの学習

一方で，識別モデル $D(\cdot)$ は，Eq. (1) を -1 倍した関数を最小化するように学習される．$D(\cdot)$ のモデルパラメータを θ_{D} とすると，θ_{D} は次式で推定される．

$$\theta_{\mathrm{D}} = \underset{\theta_{\mathrm{D}}}{\operatorname{argmin}} -V(G,D) \tag{3}$$

3　人間 GAN

人間の知覚分布を表現するために，本節では人間 GAN を提案する．人間 GAN は，**Fig. 1** に示すように，通常の GAN の識別モデルを人間の知覚評価で置換した手法である．生成モデル $G(\cdot)$ が DNN で記述され，既知の確率分布に従う乱数を入力とすることは，通常の GAN と人間 GAN で共通する．一方で，通常の GAN では DNN で記述された識別モデルを詐称して生成モデルを学習するのに対し，人間 GAN では人間による知覚評価を詐称して生成モデルを学習する．ここで，$D(\cdot)$ を人間による評価を表す知覚モデルとして再定義する．この $D(\cdot)$ は，$G(\cdot)$ から生成されたデータ $\hat{\boldsymbol{x}}_n$ を入力とし，当該入力が「知覚的にどの程度許容できるか」を 0 から 1 の値で事後確率として出力する．学習時の目的関数 $V(\cdot)$ は次式となる．

$$V(G,D) = \sum_{n=1}^{N} D\left(G\left(\boldsymbol{z}_n\right)\right) \tag{4}$$

この式が示すように，人間 GAN では実在データを使用して学習しないことに注意する．

3.1　生成モデルの学習

$G(\cdot)$ のモデルパラメータ θ_{G} は，Eq. (4) を最大化するように学習される．ここで，勾配法による反復的学習法を考える．すなわち，θ_{G} は次式で反復的に更新される．

$$\theta_{\mathrm{G}}^{(\mathrm{new})} = \theta_{\mathrm{G}} + \alpha \frac{\partial V(G,D)}{\partial \theta_{\mathrm{G}}} \tag{5}$$

ここで，α は学習係数である．第 2 項の $\partial V(G,D)/\partial\theta_{\mathrm{G}}$ は，$\partial V(G,D)/\partial\hat{\boldsymbol{x}}$ と $\partial\hat{\boldsymbol{x}}/\partial\theta_{\mathrm{G}}$ の積となる．通常の GAN では，$G(\cdot)$ と $D(\cdot)$ の両方の計算過程が微分可能であるため，この式は backpropagation を用いて実行することができる．しかしながら，人間 GAN では，$D(\cdot)$ が微分不可能であり，$\partial V(G,D)/\partial\hat{\boldsymbol{x}}$ を推定できないため，backpropagation による学習は不可能である．そこで，人間を「生成データに対して事後確率分布を出力する black-box システム」とみなした black-box 最適化手法に基づいて $\partial V(G,D)/\partial\hat{\boldsymbol{x}}$ を推定する．本研究では，摂動を用いて勾配を近似する Natural Evolution Strategies (NES) [3] を用いた学習アルゴリズムを提案する．手法の概要図を **Fig. 2** に示す．

まず，生成データ $\hat{\boldsymbol{x}}_n$ に対し，多変量正規分布 $\mathcal{N}\left(\boldsymbol{0},\sigma^2\boldsymbol{I}\right)$ からランダムに生成した摂動 $\Delta\boldsymbol{x}_n^{(r)}$ を付与する．ここで，σ は定数の標準偏差，r は摂動のインデックス $(1 \le r \le R)$，$\boldsymbol{I}$ は単位行列である．次に，摂動後の 2 つのデータ $\left\{\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}, \hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}\right\}$ を人間に提示し，それらのデータの事後確率の差分

$$\Delta D(\hat{\boldsymbol{x}}_n^{(r)}) \equiv D\left(\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}\right) - D\left(\hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}\right) \tag{6}$$

を回答させる．$\Delta D(\hat{\boldsymbol{x}}_n^{(r)})$ は -1 から 1 までの値をとる．例えば，2 つのデータのうち $\hat{\boldsymbol{x}}_n + \Delta\boldsymbol{x}_n^{(r)}$ の事後確率が非常に高い場合に人間は $\Delta D(\hat{\boldsymbol{x}}_n^{(r)}) = 1$ を返し，$\hat{\boldsymbol{x}}_n - \Delta\boldsymbol{x}_n^{(r)}$ の事後確率が非常に高い場合に人間は $\Delta D(\hat{\boldsymbol{x}}_n^{(r)}) = -1$ を返すものとする．この摂動・評価を，ある生成データ $\hat{\boldsymbol{x}}_n$ に対して R 回繰り返す．最終的には，N 個の生成データに対して上記の過程を繰り返す．

Backpropagation に用いる勾配 $\partial V(G,D)/\partial\hat{\boldsymbol{x}}$ は，次式で近似される．

$$\frac{\partial V(G,D)}{\partial\hat{\boldsymbol{x}}} = \left[\frac{\partial V(G,D)}{\partial\hat{\boldsymbol{x}}_1}, \cdots, \frac{\partial V(G,D)}{\partial\hat{\boldsymbol{x}}_n}, \cdots, \frac{\partial V(G,D)}{\partial\hat{\boldsymbol{x}}_N}\right] \tag{7}$$

$$\frac{\partial V(G,D)}{\partial\hat{\boldsymbol{x}}_n} = \frac{1}{2\sigma^2 R}\sum_{r=1}^{R}\Delta D(\hat{\boldsymbol{x}}_n^{(r)})\cdot\Delta\boldsymbol{x}_n^{(r)} \tag{8}$$

学習の各反復における各生成データに対して R 回の摂動を加えるため，最終的な問い合わせ回数は，勾配法の反復回数・生成データ数 N・摂動回数 R の積となる．

3.2　考察

本研究は，「コンピュータによって解くことが困難な課題を，人間の処理能力を利用して解決すること」であるヒューマンコンピュテーション [4] を利用したものであり，人間参加型（human-in-the-loop）の機械学習技術とみなされる．このような観点から見ると，本研究は人間を「事後確率差分を出力するシステム」とみなして，人間の知覚分布を表現する手法とみなされる．

通常の GAN は計算機のみ，提案する人間 GAN は人間のみを識別器として利用する．本研究の展望として，計算機と人間の共創，すなわち計算機と人間の両方を識別器とした生成モデル学習が考えられる．

学習データの偏り (data imbalance) は，通常の GAN に限らず学習データを用いる機械学習一般の課題である．提案する人間 GAN は，学習データに依存しないためこの課題を緩和できる可能性がある．

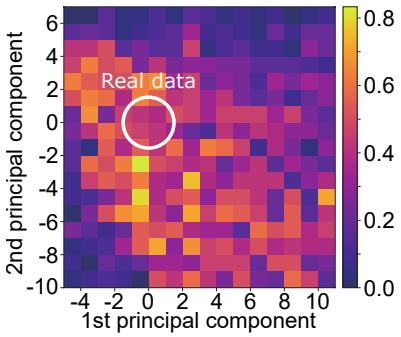

Fig. 3　主成分空間の事後確率を表すカラーマップ

3.3　課題

ここでは，人間 GAN の実装にあたり，実験的に明らかになった課題を整理する．

生成モデルの初期化　Eq. (4) に示すように，人間 GAN の生成モデルは事後確率を最大化するように学習される．故に，初期の生成モデルから生成されるデータが知覚分布の mode 近傍のみに分布する場合，学習を進めると mode collapse の問題が発生してしまう．また，生成されるデータが知覚分布から大きく離れた領域のみに分布する場合，通常の GAN と同様に gradient vanishing によりモデルパラメータが更新されない問題も発生する．これらの問題を緩和するために，後述する実験的評価では，知覚分布の領域を別途荒く推定し，その領域をカバーするような生成モデルパラメータを初期値としている．

摂動の標準偏差 σ への敏感性　σ の値の大きさに応じて，生成データに加えられる摂動の大きさが変化する．過度に小さな値の σ を用いると，人間 GAN の識別器では知覚的な差異が発生せずに勾配が消失する．一方で，過度に大きな値の σ を用いると，正確な勾配推定が行われない．後述する実験的評価では，複数の σ の値を用いて評価を行い，最終的に単一の値を使用している．

人間への問い合わせ回数　**3.1** 節で述べたように，人間への問い合わせ回数は学習反復回数・生成データ数・摂動回数の積になる．この問い合わせ回数の増加は，人間による知覚評価に係る金銭的・時間的コストを爆発させる．後述する実験的評価では，問い合わせ回数を限定するために単純な生成モデル，少量の生成データ，低次元の特徴量を使用している．

4　実験的評価

4.1　実験条件

実験的評価では，提案する人間 GAN が音声の自然性の知覚分布を表現できるかを調査する．すなわち，人間が当該音声を人間の音声として許容できる音声特徴量の知覚分布を表現する．本来，人間 GAN は実在データを利用しないが，この評価では，通常の GAN との比較，データの次元数削減，生成モデルパラメータの初期化のために，音声コーパスを利用する．使用するコーパスは，身体情報付き男・女・子どもの母音音声データベース (JVPD) [5] に含まれる女性 199 名の音声である．特徴量抽出前に音量の正規化と，16 kHz へのダウンサンプリングを行う．抽出する音声特徴量は，音声分析システム WORLD [6, 7] を用いて，5 ms 毎に抽出される 513 次元の対数スペクトル包絡である．Julius [8] による音素アライメントを行い，全時刻の音声特徴量のうち，日本語母音/a/の音声特徴量を用いる．最後に，199 名の女性の/a/の対数スペクトル包絡に対して主成分分析を施して 2 次元の主成分を得る．本研究では，多人数話者の実在データの分布は，2 次元の標準正規分布に従うと仮定し，通常の GAN と人間 GAN では，その第 1 主成分と第 2 主成分のそれぞれを平均 0，分散 1 に正規化した空間を扱う．人間による知覚評価の際には，生成された特徴量から音声波形を再合成する．まず，正規化済みの第 1, 2 主成分を生成モデルから生成し，逆正規化を行う．第 3 主成分以降と，対数スペクトル包絡以外の音声特徴量（基本周波数，非周期性指標）については，平均声話者（すなわち，第 1, 2 主成分の値が最も 0 に近い話者）の実発話の 1 フレームの特徴量を使用する．これらの特徴量は 5 ms（1 フレーム）分の音声特徴量に相当する．その後，人間による評価を容易にするために，この特徴量を 200 フレームだけコピーして得られた 1 秒分の音声特徴量から，WORLD を用いて音声波形を生成する．

4.2　実在データ分布と知覚分布の違いの確認

まず，人間 GAN の必要性を実験的に確認するために，実在データ分布と知覚分布が異なることを示す．2 次元の空間を 1.0 毎にグリッド分割し，各グリッドにおける自然性の許容度（すなわち，自然性の事後確率）を評価する．評価時には，クラウドソーシングサービス上で「人間らしくない音声であれば (1) を，人間らしい音声であれば (5) を選択してください.」というインストラクションを示し，人間に 1 つの音声の許容度を 5 段階で回答させる．ここで得られた回答の 1 から 5 のそれぞれの値を，事後確率 $D(\boldsymbol{x}_n)$ の $0.00, 0.25, 0.50, 0.75, 1.00$ に対応させる．各グリッドの事後確率の値は，複数の評価値の平均とする．評価者の数は，96 人である．

Fig. 3 に結果を示す．**4.1** 節において述べたように，実在データは平均 0，分散 1 に正規化されている．通常の GAN で表現できる分布は，この範囲に収まる．一方で，**Fig. 3** に示すように，知覚分布はこの範囲に留まらず，より広い領域に対応することがわかる．この広い領域は，通常の GAN では表現することが出来ないため，人間 GAN の枠組みによる知覚分布の表現が必要である．

4.3　生成モデル学習による生成データの変化

次に，生成モデル学習の様子を定性的に確認するために，学習の各反復における生成データをプロットする．生成モデルに入力する乱数は，2 次元の一様分布 $U(-1, 1)$ に従うものとする．この乱数は，学習のはじめにランダムに生成され，学習中は固定される．生成モデルは，2 ユニットの入力層，2×4 ユニット

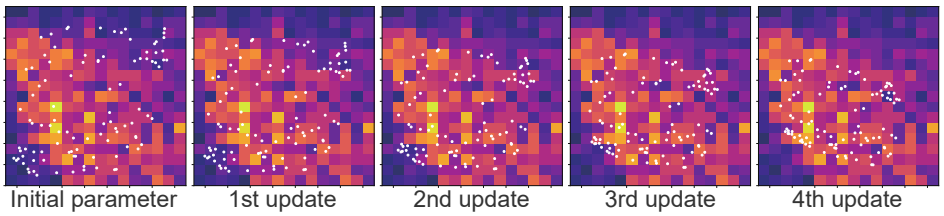

Fig. 4　各学習回数における生成データの変化

の sigmoid 隠れ層，2 ユニットの線形出力層を持つ Feed-Forward neural network である．初期の生成モデルパラメータはランダムに決定する．ただし，**Fig. 3** における事後確率の高い領域をカバーするような初期生成データが得られるまで，初期モデルパラメータのランダム生成を繰り返す．最適化アルゴリズムとして学習率 $\alpha = 0.0015$ の確率的勾配降下法を用いる．生成する音声特徴量数 N は 100，摂動回数 R は 5，学習の反復回数は 4 とする．NES の標準偏差 σ は 1.0 とする．人間による評価には，クラウドソーシングサービスを使用する．評価時には，クラウドソーシングサービス上で「2 つの音声を聴いて，1 個目の方が人間らしいほど (1) を，2 個目の方が人間らしいほど (5) を選択してください．両方同程度なら (3) を選択してください.」というインストラクションを示し，人間に 2 つの音声の許容度の差異を 5 段階で回答させる．ここで得られた回答の 1 から 5 の値を，それぞれ事後確率の差分 $\Delta D(\hat{\boldsymbol{x}}_n^{(r)})$ の $1.0, 0.5, 0.0, -0.5, -1.0$ に対応させる．

Fig. 4 に学習の様子を示す．各点は生成データである．可視化のため，**Fig. 3** の事後確率を重ねて表示しているが，学習そのものに **Fig. 3** の結果は使用していないことに注意する．これらの図より，学習の反復により，生成データは事後確率の高い領域に遷移していることがわかる．故に，人間 GAN は，知覚分布を表現するような生成モデル学習を可能にすることが定性的に確認できる．

4.4　生成モデル学習による事後確率の増加

最後に，学習の反復により事後確率が増加することを定量的に確認する．ここでは，学習に用いた乱数とは別の乱数を発生させ，新たな特徴量を生成する．以降では，学習に使用した乱数で生成した特徴量を closed データ，学習に使用していない乱数で生成した特徴量を open データと呼ぶ．Open データに対する事後確率は，**4.2** 節と同様に，クラウドソーシングサービス上で人間に評価させる．

Fig. 5 に，各反復における事後確率の箱ひげ図を示す．この図から，学習の反復により，closed データのみならず，open データに対しても事後確率の増加が見られる．故に，人間 GAN により学習された生成モデルが知覚分布を表現しうることを定量的に確認できる．

5　まとめ

本研究では，人間の許容できるデータの範囲（知覚分布）を表現可能な生成モデル学習のために，人間による知覚評価を詐称する敵対的生成ネットワーク

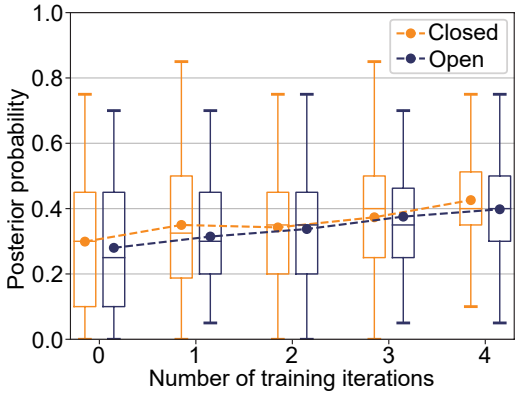

Fig. 5　各学習回数における事後確率

を提案した．その有効性を示すために音声の自然性に関する実験的評価を行い，提案法が知覚分布を表現する生成モデルを学習することを示した．今後は，人間と計算機の共創に基づく敵対的生成ネットワークを検討する．

謝辞：本研究の一部は，セコム科学技術振興財団の支援を受けて実施した．また，本研究開発は総務省 SCOPE(受付番号 182103104) の委託を受けた．

参考文献

[1] I. Goodfellow et al., “Generative adversarial nets,” in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[2] Y. Saito et al., “Statistical parametric speech synthesis incorporating generative adversarial networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[3] A. Ilyas et al., “Black-box adversarial attacks with limited queries and information,” in *Proc. ICML*, vol. 2, Stockholm, Sweden, Jul. 2018, pp. 2137–2146.

[4] A. J. Quinn, B. Bederson, “Human computation: a survey and taxonomy of a growing field,” in *CHI ’11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, May 2011, pp. 1403–1402.

[5] “Vowel database: Five Japanese vowels of males, females, and children along with relevant physical data (JVPD),” http://research.nii.ac.jp/src/en/JVPD.html.

[6] M. Morise et al., “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[7] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, Nov. 2016.

[8] A. Lee et al., “Julius — an open source real-time large vocabulary recognition engine,” in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.