一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2018-132,SIP2018-138,SP2018-94(2019-03)

アノテータのラベル付与能力を考慮した電話応対音声の好感度推定モデル学習法の検討

神山歩相名† 安藤 厚志† 増村 亮† 小橋川 哲† 青野 裕司†

† NTT メディアインテリジェンス研究所 〒239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †kamiyama.hosana@lab.ntt.co.jp

あらまし 本研究では，コンタクトセンタのオペレータの応対の好感度の自動推定に取り組む．従来の好感度推定では，ラベルが対象音声データ・アノテータによって異なることがあるため，各対象音声データごと多数のアノテータでラベル付けを行い，好感度推定モデル学習用の精度の高い正解ラベルを得ていた．しかし，コンタクトセンタのオペレータの応対音声は，長時間の音声のため 1 音声に付与できるラベルの数は限られ，ラベル付与誤りにより精度の低い正解ラベルとなる．そこで本研究では，1 つの音声辺り少数のアノテータのラベルを用いて，アノテータの付与能力を考慮した高精度な好感度推定モデルを学習する手法について提案する．本研究では，2 つのアプローチによる好感度推定モデル学習法を提案する．1 つ目は，観測ラベルからアノテータの影響を排除したラベルを推定を行う．観測ラベルにはラベル付与誤りを含むため，アノテータのラベルの誤りやすさを考慮した好感度ラベルを推測し，学習に用いる．2 つ目は，本来の好感度からアノテータの能力に応じてラベル付与の誤りが発生する過程をニューラルネットワークに内包することで，ラベル付与誤りを考慮したモデル学習を実現する．実験の結果，従来の好感度推定モデル学習法から最大誤り削減率 12.0% を達成した．

キーワード 好感度, コンタクトセンタ オペレータ, アノテータ, ニューラルネットワーク

Likability Estimation Model Training of Call-center Agents Based on Annotators' Skills

Hosana KAMIYAMA†, Atsushi ANDO†, Ryo MASUMURA†, Satoshi KOBASHIKAWA†, and Yushi AONO†

† NTT Media Intelligent Laboratories, NTT Corporation 1–1 Hikarinooka, Yokosuka-shi, Kanagawa, 239–0847 Japan
E-mail: †kamiyama.hosana@lab.ntt.co.jp

Abstract This paper proposes a new technique for estimating the likability of call-center agents. Most techniques of likability estimation collect many annotations per sample since the annotations often include the variability of annotator evaluations. Due to the inability to collect sufficient annotations, which are labeled by experts such as supervisor of call-center, from actual call-centers, the few annotations per call available cannot eliminate the variability. We proposes two likability model training techniques from the few annotations per call based on annotators' skill. First, our technique provides reasonable labels, which eliminated from the influence of annotators' skills, for model training from observed labels. We also proposes a new neural network architecture that has a layer considering the variability of observed labels from each annotator. Our proposal achieved 12.0% error reduction compared with baseline techniques.

Key words Likability, Call-center agent, Annotator, Neural Network

1. はじめに

コンタクトセンタでは，オペレータの電話応対の品質の評価を行っている．電話応対の品質評価は，一般的には，スーパーバイザがオペレータの通話をランダムにサンプリングして通話を聞き，評価を行っている．全ての通話の評価は難しいため，

近年では言語的なマッチングを行っている電話応対の自動評価するシステムが提案されている [1], [2].（例えば，「本人確認ができているか？」「クロージングの確認事項は話したか？」等）．一方，電話応対では「オペレータが感じの良い応対か？」といった，応対の好感度もオペレータの評価指標として重要である．そのため，オペレータの電話応対の好感度を直接自動推定する技術が必要とされている．

音響情報から，発話者の好感度を推定する手法は数多く提案されている．Interspeech 2012 Speaker Trait Challenge [3], [4] では，話者の好感度を推定するためのデータベース “Speaker Likability Database (SLD)” [5] が提供され，主に特徴量とモデル化を中心に研究が取り組まれている [6]〜[9]．クラウドソーシングで収集した好感度のデータベース [10] を用いて，好感度を推定する研究も取り組まれている [11].

従来技術の課題として，精度の高い学習ラベルを得るために，多数のアノテータが必要である点があげられる．これは好感度のような主観的なラベルは，同じ音声でもアノテータによって付与されるラベルが異なることがあり，少数のアノテーションではラベル付与誤りを含む可能性があるためである．例えば，SLD [5] では 1 つの音声データ（平均時間長 3.2 秒）あたり 32 名のアノテータがラベル付与を行っている．また，文献 [10] でも同程度の数秒の音声に対して，1 音声辺り 29 名にてアノテーションを行っている．これらは高精度なラベルとして，多数のアノテータが一致しているデータを採用している．しかし，コンタクトセンタのオペレータ応対音声は，コンタクトセンタのスーパーバイザ等の有スキル者が通話を全体（数分〜十数分）通じてアノテーションをする必要があり，短時間の音声よりラベル付与よりコスト高く，1 音声データ辺り多人数でラベル付与をすることは現実的ではない．特に電話応対音声は，短時間の音声に比べて音声の時間長が長く，好感度を正しく判定するのは短時間の音声よりも難易度が高く，多数のアノテータのラベルが一致する学習データを十分な数得るのは難しい．そのため，電話応対では少数のアノテーションからでも高精度な好感度推定モデルを学習することが求められている．

そこで，1 つの音声データの少数のラベルから高精度な好感度推定モデルを学習する手法について提案する．本研究では，ラベルの不一致はアノテータの能力不足が原因により生じると仮定をおき（これを以降「ラベルのブレ」とする），次の 2 つのアプローチからモデル学習行う．

（1） アノテータの能力を考慮した学習ラベルの生成

（2） アノテータの能力の影響を考慮する機構を内包した深層学習モデル

(1) において従来技術では用いていなかった，ブレのある学習データを用いて精度の向上を図る．また，(2) にもアノテータの能力の影響によるブレを考慮したモデル化を行うことで精度の向上を図る．これらの手法は，ラベルの数が少数であってもアノテータの能力の影響を排除することで，精度の向上が図れると考えられる．(1) のラベル生成おいては, アノテータのラベルはアノテータの能力によって誤ったラベルが付与されていると考え，アノテータの能力に依存しない学習ラベルを生成する．本研究では，次の 3 つのアプローチの手法を本好感度推定のタスクに適用する．

(1a) ブレは全アノテータで等しく起こると仮定して学習ラベルを生成（Soft-label [12], [13]）

(1b) ブレはアノテータごとに異なり，能力の高いアノテータのラベルを重視して学習ラベルを生成（Evaluator Weighted Estimator(EWE) [14]）

(1c) ブレはアノテータごとに異なり，アノテータごと確率的に発生すると仮定し，確率的に期待される学習ラベルを生成（潜在変数モデル [15]）

また，(2) の深層学習モデルでは，アノテータの付与した好感度ラベルは，本来の好感度からアノテータの能力によりブレが生じることを確率的に表現したネットワークを構築する．(1) のアプローチの最も最適な好感度推定モデルから，アノテータの能力を考慮した機構を構築して，モデルを更新することで更なる精度向上を図る．

以下，2 節にて本研究で取り組む好感度推定モデルの枠組みと，従来手法の学習ラベルの与え方について説明する．3 節では，提案手法のラベル推定法，および好感度推定モデル学習法について述べる．4 節では，電話応対音声のデータセットを用いて評価を行い，提案手法の評価を行う．5 節では，本稿のまとめを述べる．

2. 従 来 手 法

本節では，従来の好感度推定手法について述べる．

2.1 問 題 設 定

本研究では，コンタクトセンタのオペレータの好感度の推定に取り組む．ある音声 i の好感度の推定を識別モデルによって捉える場合，次のように定式化ができる．

$$\hat{y}_i = \arg\max_{y_i \in L} p(y_i|\mathbf{X}_i, \mathbf{\Theta}) \tag{1}$$

$\hat{y}_i$ は音声 i の好感度クラスの推定値，$\mathbf{X}_i = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_{n_i}})$ は音声 i の特徴量系列，L は好感度のクラスの集合，$p(y_i|\mathbf{X}_i, \mathbf{\Theta})$ は好感度推定モデルにより求めた好感度クラスごとの事後確率であり，$\mathbf{\Theta}$ は好感度推定モデルのパラメータである．本研究は，この好感度推定モデルの推定精度を高めるためにモデルパラメータ $\mathbf{\Theta}$ をアノテータのラベルから最適化する問題となる．

本研究では好感度を 2 値（*likable*, *non-likable*）のクラスで推定する問題設定とし，音声 i が好感度が *likable* のときは $y_i = 1$, 好感度が *non-likable* のときは $y_i = 0$ と正解ラベルと定義する．また，本研究では好感度推定モデル $p(y_i|\mathbf{X}_i, \mathbf{\Theta})$ に時系列データの識別で用いられている注意機構付き Long-Short Term Memory (LSTM-Attention) を用いる [12]．LSTM-Attention に基づく好感度推定モデルのネットワーク構造を，図 1 に示す．このネットワークは，LSTM 層に音声の特徴量系列 $\mathbf{X}_i = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_{n_i}})$ を入力し，LSTM 層の出力を注意機構に入力する．注意機構の活性化関数にソフトマックス関数を用いることで，好感度は各クラスの確率値として出力される．

2.2 モデルの学習

好感度推定モデルの学習は始めに各音声に対して複数名のアノテータがラベル付与を行い，複数のラベルから学習データの正解ラベルを決定する．続いて正解ラベルと音声の特徴量を用いて，好感度推定モデルを学習する．定式化すると次のようになる．始めに音声 i について，アノテータ j が付与したラベル

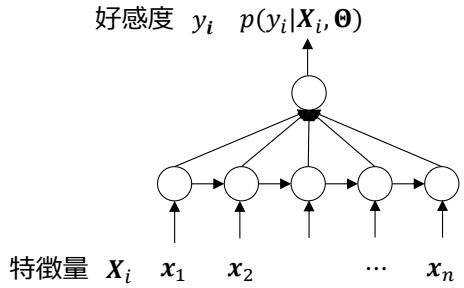


図 1　好感度推定モデルのネットワーク構造

を $y_i^{(j)}$ とすると正解ラベル y_i の決定は次のようになる．

$$y_i = f(y_i^{(j)}|j \in J_i) \tag{2}$$

J_i は音声 i のラベルを付与しているアノテータの集合，f は複数のアノテータラベル $y_i^{(j)}$ から，正解ラベル y_i を決定する関数とする．学習は，好感度推定モデル $p(y_i|\mathbf{X}_i,\boldsymbol{\Theta})$ について Cross-entropy 誤差に基づき下記損失関数 E を小さくするようにモデルパラメータ Θ を更新する．

$$E = -\sum_i \sum_c y_{i,c} \log p(y_i = c|\mathbf{X}_i,\boldsymbol{\Theta}) \tag{3}$$

c はクラス番号であり $c=1$ のとき *likable* のクラス，$c=0$ のとき *non-likable* のクラスの番号を示す．$y_{i,c}$ は正解ラベル y_i を one-hot 表現した際の c 番目のベクトルの要素である．例えば，正解ラベル $y_i=1$ の場合，ベクトルで $(y_{i,0}, y_{i,1}) = (0,1)$ となり，正解ラベル $y_i=0$ の場合，$(y_{i,0}, y_{i,1}) = (1,0)$ となる．

2.3　従来手法のラベルの決定

好感度は同じ音声でもアノテータによってラベルが異なることがある．このとき一般的には，全会一致しているデータを選定して学習データとし，次のように定式化ができる．

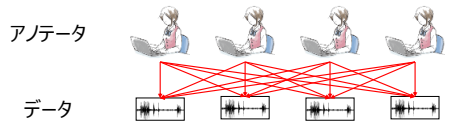
$$y_i = f(y_i^{(j)}|j \in J_i) \tag{4}$$

$$= \begin{cases} 0 & (\text{if } \forall j \in J_i, y_i^{(j)} = 0) \\ 1 & (\text{if } \forall j \in J_i, y_i^{(j)} = 1) \\ \text{学習から除外} & (\text{otherwise}) \end{cases} \tag{5}$$

$f(y_i^{(j)}|j \in J_i)$ は，ラベル決定をする関数である．従来手法は，全てのアノテータが同じではないラベルの時は学習データから除外する．

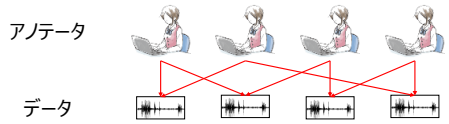
2.4　従来手法の課題

従来手法では，1 つの音声データ辺り少数のアノテータによるラベルから精度の高い学習の実現が課題となる．従来は，短時間の音声のため図 2 (a) のように全音声に対して全アノテータがラベルを付与しているが，コンタクトセンタのオペレータの好感度には 1 音声辺りに付与できるアノテーション数は限られ，図 2 (b) のように 1 つの音声データあたり少数のアノテータのラベルとなる．好感度のアノテーションはアノテータによってラベルの不一致があることが，従来技術でも報告されている [5]．このラベルが不一致のアノテータのラベル付与の誤りとすると，従来技術と同様に全会一致したラベルを，1 音声辺り限られた数からのラベルを採用する場合，正解ラベルが誤った学習データが混在することとなる．その結果，従来手法の学習データで学習した場合は，好感度推定モデルは十分な推定精度が得られないこととなる．



データi / アノテータj	1	2	3	4
1	0	1	0	1
2	0	1	0	0
3	0	1	1	1
4	0	1	1	0
決定ラベル	0	1	-	-

(a) 理想的なラベル付与



データi / アノテータj	1	2	3	4
1	0	1	-	-
2	0	-	-	0
3	-	1	1	-
4	-	-	1	0
決定ラベル	0	1	1	0

(b) 実際のコンタクトセンタで行われるラベル付与

図 2　理想的なラベル付与と実際に行われるラベル付与

3.　提 案 手 法

本節では，1 音声のアノテーション数が限られている状況下で，好感度推定モデルを学習する手法について説明する．本節では，2 つのアプローチによるモデル学習法について説明し，1 つ目はブレを考慮した学習ラベルを生成する手法，2 つ目はアノテータのブレを考慮した機構を内包したネットワークについて説明する．

3.1　学習ラベルの生成手法

本研究は，ラベルのブレはアノテータのラベル付与の能力不足に起因すると仮定した．これはアノテータがラベルの付与に慣れていない場合（例えば，アノテータが新たにスーパーバイザになった人物のとき等）には，正しく音声に対してラベルの付与ができず，誤ったラベルが付与されると考えた．

従来技術では，多人数で全会一致したラベルを学習データとするため，ラベル付与が難しい音声は学習データには含まれない前提となっている．しかし，本研究では 1 データ辺りのアノテーション数が少ないため誤りを含むことがある．また，従来技術はアノテータのラベル付与の能力についてまでは考慮がされていなかった．

そこで，アノテータ能力を考慮して緩やかに各クラスの「クラスらしさ」を情報としてラベルを与える学習ラベルの生成手法が好感度以外のタスクで提案されている [12], [14], [15]．本研究では，好感度の推定モデルの学習に，これら学習ラベルの生成手法を適用する．

3.1.1　Soft-label

Soft-label は，アノテータのラベル付与の能力が等価と考え，

複数のブレを含むラベルから緩やかにクラスらしさを表現したラベルを与えて学習する [12]．Soft-label は，感情認識等で用いられており，音声がアノテータによってラベルがブレた場合，各クラスに付与されたラベルの割合で目的関数を与える．このときラベルを決定する関数 f_{soft} は下記のようになる．

$$y_{i,c} = f_{\text{soft}}(y_i^{(j)}|j \in J_i))_c \tag{6}$$

$$= \frac{\sum_{j \in J_i} y_{i,c}}{|J_i|} \tag{7}$$

$|J_i|$ は音声 i をラベル付与したアノテータの数を示す．また，$f_{\text{soft}}(y_i^{(j)}|j \in J_i)_c$ は，学習ラベルを one-hot 表現を行った c 番目の要素となる．例えば，アノテータ数が 3 名存在し，各アノテータが $(y_i^{(1)}, y_i^{(2)}, y_i^{(3)}) = (0, 0, 1)$(*non-likable*,*non-likable*,*likable*) とラベル付与をしたとき，各クラスの目的関数 $(y_{i,0}, y_{i,1}) = (2/3, 1/3)$ を与えて学習を行う．

3.1.2 Evaluator Weighted Estimator (EWE)

EWE は，アノテータごとにラベル付与の能力が異なると考え，能力の高いアノテータの重視して緩やかにクラスらしさを表現したラベルを与えて学習する [14]．EWE はアノテータの能力を評価の平均らしさと仮定する．アノテータが評価したラベルと，全アノテータが評価したラベルの平均点との相関係数を能力値として求め，相関係数が大きいアノテータを重みづけした Soft-label の手法と位置付けられる．このときラベルを決定する関数 f_{EWE} は下記のようになる．

$$y_{i,c} = f_{\text{EWE}}(y_i^{(j)}|j \in J_i)_c \tag{8}$$

$$= \frac{\sum_{j \in J_i} r_j y_i^{(j)}}{\sum_{j \in J_i} r_j} \tag{9}$$

r_j は，アノテータ j と全アノテータが評価したラベルとの相関係数である．相関係数 r_j は，下記式により求めることができる．

$$r_j = \frac{\sum_i \left(y_i^{(j)} - E_i[y_i^{(j)}]\right)\left(E_j[y_i^{(j)}] - E_{i,j}[y_i^{(j)}]\right)}{\sqrt{\sum_i \left(y_i^{(j)} - E_i[y_i^{(j)}]\right)^2}\sqrt{\sum_i \left(E_j[y_i^{(j)}] - E_{i,j}[y_i^{(j)}]\right)^2}} \tag{10}$$

$E_v[\cdot]$ は変数 v についての平均を示し，$E_i[y_i^{(j)}]$ はアノテータ j が全データ i に対してラベル付与したラベルの平均値，$E_j[y_i^{(j)}]$ は各データ i に対してアノテータ j がラベルを付与した平均値，$E_i, j[y_i^{(j)}]$ は全ラベルデータの平均値となる．なお，文献 [14] では，負の相関が認められるアノテータについてはノイズとして学習データから除外を行う（$r_j < 0$ の場合，$r_j = 0$ とする）．例えば，アノテータ数が 3 名存在し，各アノテータが $(y_i^{(1)}, y_i^{(2)}, y_i^{(3)}) = (0, 0, 1)$(*non-likable*,*non-likable*,*likable*) とラベル付与をしたとき，各アノテータの相関係数が $(r_1, r_2, r_3) = (0.75, 0.0, 0.25)$ の場合，各クラスの目的関数 $(y_{i,0}, y_{i,1}) = (0.75, 0.25)$ となり，$j = 1$ のアノテータが重視されたラベルとなる．

3.1.3 潜在変数モデルに基づくラベル推定

潜在変数モデルに基づく手法は，アノテータごとにラベル付与の能力が異なると考えるが，正しい/誤ったラベルはアノテータごとに確率的に発生すると考えて，確率的に期待されるラベルを求めて学習する [15]．Soft-label，EWE は，各音声に閉じて重みによりラベルを生成していたが，1 音声辺りのアノテータ数が限られるため，誤って付与したラベルも含まれる問題があった．各音声に閉じずに，確率的に尤もらしいラベルの期待値を求めることで，よりアノテータの能力の影響を排除したラベルが推定できることが期待できる．本手法に基づく潜在変数モデルを図 3 に示す．本手法は，音声 i には真の好感度 h_i の潜在変数が存在し，アノテータが正しくラベル付与できる確率 α_j, β_j に基づきアノテーションラベル $y_i^{(j)}$ が生成されると仮定する．この生成モデルでは，次のようにラベルの生成過程を Bernoulli 分布でモデル化を行う．

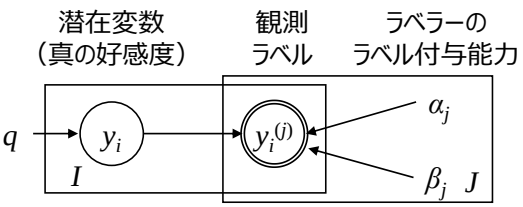


図 3　潜在変数モデル

$$p(h_i|q) = \text{Bernoulli}(h_i|q) \tag{11}$$

$$p(y_i^{(j)}|\alpha_j, \beta_j, h_i = 1) = \text{Bernoulli}(y_i^{(j)}|\alpha_j) \tag{12}$$

$$p(y_i^{(j)}|\alpha_j, \beta_j, h_i = 0) = \text{Bernoulli}(y_i^{(j)}|\beta_j) \tag{13}$$

ベルヌーイ分布は，$\text{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{1-x}$ にて示される 2 値の確率分布である．上記を仮定の上，EM アルゴリズム等でパラメータ q, α_j および β_j を，観測ラベル $y_i^{(j)}$ から求める．パラメータ推定後，真の好感度の期待値 $E_{q,j}[h_i]$ を求め目的関数とする．このときラベルを決定する関数 f_H は下記のようになる．

$$y_{i,c} = f_H(y_i^{(j)}|j \in J_i)_c = E_{q,j}[h_i = c] \tag{14}$$

$$= \frac{q_c \prod_j p(y_i^{(j)}|\alpha_j, h_i = c)}{\sum_{c'} q_{c'} \prod_j p(y_i^{(j)}|\alpha_j, \beta_j, h_i = c')} \tag{15}$$

q_c は，パラメータ q に基づく値で $(q_0, q_1) = (1-q, q)$ である．本手法は，アノテータ j の能力に応じて真の好感度を推定するため，各音声に閉じたラベルにならず，誤りやすいアノテータが付与したラベルは誤りやすさを考慮した，緩やかなクラス表現の学習ラベルとなることが期待できる．

3.2　ラベルのブレやすさを考慮したネットワーク

2 つ目のアプローチは，学習ラベルの生成ではなく，アノテーションのブレをモデルに内包し，観測ラベルからブレを考慮して好感度推定モデルを学習する手法である．提案するネットワークの構造を図 3.2 に示す．本ネットワークは，これまでの好感度推定のネットワークの出力層 $p(y_i|\mathbf{X}_i, \boldsymbol{\Theta})$ とアノテータの能力 $p(y_i^(j)|y_i, j, \alpha, \beta)$ に基づき，観測ラベル $y_i^{(j)}$ が出現する確率を解析的に与えたネットワークとなる．アノテータ j が真の好感度 y_i に対してラベル $y_i^{(j)}$ を付与する確率 $p(y_i^{(j)}|y_i, j, \alpha, \beta)$ を用いて下記のようになる．

$$p(y_i^{(j)}|\mathbf{X}_i, j, \alpha, \beta, \boldsymbol{\Theta}) = \sum_{y_i} p(y_i^{(j)}|y_i, j, \alpha, \beta) p(y_i|\mathbf{X}_i, \boldsymbol{\Theta}) \tag{16}$$

$p(y_i^{(j)}|y_i, j, \alpha, \beta)$ は，アノテータの離散番号 j を連続ベクトルに変換する Embedding 層及び Softmax 関数により，アノテー

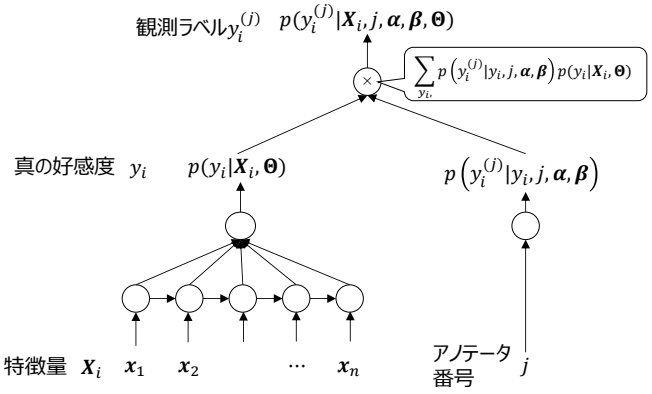


図 4　提案するネットワーク

タ j おいて各好感度パラメータの確率値を出力する層を実現することができる．α, β は，正しくラベルを出力する確率値を示す学習可能なモデルパラメータを示す．このネットワークは，観測ラベルからモデルパラメータ $\boldsymbol{\Theta}, \alpha, \beta$ を学習を行い，下記のような損失関数となる．

$$E = -\sum_i \sum_c y_{i,c}^{(j)} \log \sum_{y_i} p(y_i^{(j)}|y_i, j, \alpha, \beta) p(|\mathbf{X}_i, \boldsymbol{\Theta}) \tag{17}$$

$y_{i,c}^{(j)}$ は，観測ラベルの one-hot 表現したときのベクトルの c 番目の要素である．このネットワークでは，好感度推定にはアノテータの能力に関係のない好感度推定モデル $p(y_i|\mathbf{X}_i, \boldsymbol{\Theta})$ 部分を用いて推定する．アノテータの能力の影響を排除したモデルが学習できると考えられる．

4. 実　　験

提案手法の有効性を確認するため，評価実験を行った．

4.1 データセット

本稿では，コンタクトセンタの電話応対技能試験のデータセットを用いる．電話応対技能試験は，オペレーターの電話応対品質（発声スキル，印象（好感度を含む），コミュニケーションスキル等）を評価する試験である．電話応対のタスクが予め決めらた上で，試験官が顧客役，受験者がオペレータ役の 2 名の通話が収録されており，それぞれの通話について好感度を含めた評価ラベルが付与されている．

4.1.1 電話応対技能試験データセット

試験は，はじめに試験官・受験者共にタスクの状況設定を読ませる（例えば，当該タスクの会社のサービス概要，予約状況等）．試験官は，さらに予め定められた顧客のシナリオを読むことができる（例えば，商品の発送遅れに対するクレーム，タクシー配車の日時希望 等）．試験が開始されると，顧客は予め設定されたシナリオを演じる．オペレータ役である受験者は，予め顧客のシナリオを把握できていないため，顧客の状況をその場で判断し応対を行う．1 通話は，平均 2.9 分であり，家電・保険・タクシー・ホテル・飲食店・化粧品・健康食品販売・出版社等の販売受付，予約，故障受付，問い合わせ，インバウンド営業，アウトバウンド営業等が収録されている．サンプリング周波数は 8kHz で，試験官および受験者はモノラルで収録されている．通話はモノラルで収録されているため，一部の音声（1174 通話）についてはオーバーラップ区間を除いて試験官および受験者それぞれの発話区間を抽出してある．

表 1　電話応対技能検定データセット

	アノテータ数	ラベル一致数	データ量
2 名ラベルデータ	2	2	1174
4 名ラベルデータ	4	2	629
		3	423
		4	340

4.1.2 ラベル付与

電話応対技能試験データセットでは，それぞれの通話に，試験官とは異なるアノテータ 2 名または 4 名が応対好感度について 5 段階で好感度 (好感度 高 5〜低 1) を評価を行ってある．アノテータは計 245 名が評価を行っており，分担してラベル付与を行っている．この元のデータセットのラベルは，90% のラベルが 3 または 4 のラベルが付与されている．そこで，今回は 3 以下を好感度 低，4 以上を好感度 高の 2 値のラベルに変換した．

2 名以上ラベルが付与されており，2 名のラベルが一致したステレオ通話は 1,174 通話得られた．また，4 名ラベルを付与した通話は 629 通話となり，4 名全員のラベルが一致した通話は 340 通話となった．各データセットについて表 1 に示す．

4.2 実 験 条 件

好感度推定の特徴量は，Interspeech Speaker Trait Challenge [3], [4] の Baseline の特徴量セットで用いられている 6125 次元の特徴量を，オペレータの発話区間ごと OpenSMILE [16] にて抽出し，事前実験によって特徴量選択を行った [17]．好感度推定するモデルには LSTM1 層の LSTM-Attention を用いて，発話単位の Z 正規化した特徴量を入力した．隠れ層のユニット数は 32 で，パラメータ更新には Adam で学習率は 0.0001，Dropout は 0.5 とした．

電話応対技試験データセットの全ラベルデータを用いてアノテータの能力に関わるパラメータ（EWE に用いる相関係数，潜在変数モデルにおけるアノテータの能力のパラメータおよび事前分布パラメータ）と真の好感度の期待値を求めた．続いて，4 名一致している 340 通話を 10 分割を行い，1/10（34 通話）を共通の評価セットとして 10 分割交差検証で評価を行った．学習はモデルの初期値パラメータを変えて 5 回実施し，最も高い精度において比較を行った．評価は，表 1 の各データ量における各ラベル生成の手法の推定精度を比較した．

比較は，全会一致したラベルで好感度モデルを学習した手法を Baseline として，生成したラベルを用いた手法および，生成したラベルから構築した深層学習モデルの手法を比較した．また最大 4 名のアノテーションされているデータを用いて，各手法について一致しているラベル数に関する精度について比較を行った．

4.3 結　　果

各学習データの好感度の推定結果を表 2 に示す．実験の結果，提案手法を併用して 2 名のアノテータ（1174 通話）で学習した結果が最も高い精度を達成した．2 名のアノテータの Baseline と比較すると，最大 12.0% の誤り削減率を達成した．

Baseline については，2 名が一致したデータセット 1174 通

表 2　好感度推定実験結果

ラベル一致数 (アノテーション数)		能力の考慮	2 名 (2 名)	2 名 (4 名)	3 名 (4 名)	4 名 (4 名)
Baseline		なし	0.779	0.753	0.768	0.782
ラベル生成	Soft-label	重み (等価)	0.779	0.781	0.785	0.779
	EWE	重み (非等価)	0.779	0.775	0.782	0.774
	潜在変数	確率	0.788	0.782	0.775	0.776
深層学習モデル		モデル内包	0.806	0.797	0.776	0.776

話と，4 名がアノテーションを行った 629 通話を比較すると，データ量が多い 1174 通話のほうが精度が高い．これは，データ量が少なくなったこととが精度が低下したものと考えられる．しかしアノテーション数を増やしたところ，推定精度は向上した．これは，ラベルの精度が高まったためと考えられる．

ラベル生成の手法では，アノテータ数が 3 名以下のときは Baseline より高い推定精度であった．よって，提案手法がアノテータ数が少ない時の有効性を確認した．また，2 名が一致したラベルを用いているときは潜在変数モデルが最も推定精度が良かった．これは，潜在変数モデルは他音声のラベルも用いて学習ラベルを生成してしているため，「そのクラスらしさ」が Soft-label，EWE より表現できていたためと考えられる．一方，3 名以上の場合は Soft-label，EWE が潜在変数モデルより精度が同程度以上となり，Soft-label,EWE ともに学習に与える「そのクラスらしさ」を表現できていると考えられる．1 音声 2 名以下のアノテーション数等，ラベル付与が難しい状況下で潜在変数モデルは有効であると言える．

さらに，提案したアノテータのブレを深層学習モデルでは，潜在変数モデルよりさらに良い精度が得られた．よってブレを考慮するネットワークについても有効性が確認された．

5. ま と め

本稿では，コンタクトセンタのオペレータの好感度推定に向けたモデルの学習手法について提案した．コンタクトセンタでは，1 通話あたり少人数によるラベル付けとなるため，ラベルのブレが学習データの精度低下に影響を与えていた．本研究では，アノテータの能力を排除した学習データの生成，およびアノテータの能力を考慮した深層学習モデルにより高精度な好感度推定モデル学習法を提案した．ラベルの生成では，少数の観測ラベルからアノテータの能力の影響を排除し，「そのクラスらしさ」を緩やかに学習するラベルを生成する．また，深層学習モデルでは，アノテータの能力に基づくラベルのブレやすさをニューラルネットワークに内包することで，観測ラベルから直接好感度の推定モデルの学習を行う．実験により，ラベル生成の手法では 1 音声 3 名以下の少数のアノテーション数であれば，従来手法より高い精度で推定し，潜在変数モデルの手法が最も良い精度であった．また，深層学習モデルも，潜在変数モデルからさらに精度が好感度推定モデルが学習できることが確認された．1 音声あたり 2 名の少数のアノテーション数で比較すると，最大 12.0% の誤り削減率を達成した．

文　献

[1] G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury, “Automated quality monitoring for call centers using speech and nlp technologies,” in *Proc. NAACL HLT*, 2006, pp. 292–295.

[2] S. Roy, R. Mariappan, S. Dandapat, S. Sricastave, S. Galhotra, and B. Peddamuthu, “Qart: A system for real-time hoilistic quality assurance for contact center dialogues,” in *Proc. AAAI*, 2016, pp. 3768–3775.

[3] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, “The interspeech 2012 speaker trait challenge,” in *Proc. Interspeech*, 2012, pp. 254–257.

[4] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, R. v. Son F. Burkhardt, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, “A survey on perceived speaker traits: Personality and likability and pathology and and the first challenge,” in *Computer Speech and Language*, 2015, vol. 29, pp. 100–131.

[5] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ““would you buy a car from me?” – on the likability of telephone voices,” in *Proc. Interspeech*, 2011, pp. 1557–1560.

[6] J. Pohjalainen, S. Kadioglu, and O. Räsänen, “Feature selection for speaker traits,” in *Proc. of Interspeech*, 2012, pp. 270–273.

[7] D. Wu, “Genetic algorithm based feature selection for speaker trait classification,” in *Proc. of Interspeech*, 2012, pp. 294–297.

[8] H. Buisman and E. Postma, “The log-gabor method: Speech classification using spectrogram image analysis,” in *Proc. of Interspeech*, 2012, pp. 518–521.

[9] C. Montacié and M. Carat, “Pitch and intonation contribution to speakers’ traits classification,” in *Proc. of Interspeech*, 2012, pp. 526–529.

[10] L. F. Gallardo, R. Z. Jiménez, and S. Möller, “Perceptual ratings of voice likability collected through in-lab listening tests vs. mobile-based crowdsourcing,” in *Proc. Interspeech*, 2017, pp. 2233–2237.

[11] S. Hantke, E. Marchi, and B. Schuller, “Introducing the weighted trustability evaluator for crowdsoucing exemplified by speaker likability classification,” in *Proc. LREC*, 2016, pp. 2156–2161.

[12] H. M. Fayek, M. Lech, , and L. Cavedon, “Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels,” in *Proc. IJCNN*, 2016, pp. 566–570.

[13] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, “Of all things the measure is man: Automatic classification of emotions and inter-labeler consistency,” in *Proc. ICASSP*, 2005, pp. 317–320.

[14] M. Grimm and K. Kroschel, “Evaluation of natural emotions using self assessment manikins,” in *Proc. ASRU*, 2005, pp. 381–385.

[15] A. P. Dawid and A. M. Skene, “Maximum likelihood estimation of observer error-rates using the em algorithm,” in *Journal of the Royal Statstical Society. Series C*, 1979, vol. 28, pp. 20–28.

[16] F. Eyben, M. Wöllmer, and B. Schuller, “opensmile - the munich versatile and fast open-source audio feature extractor,” in *Proc. of ACM Multimedia*, 2010, pp. 1459–1462.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, and R. Weiss, “Scikit-learn: Machine learningin python,” in *Journal of Machine Learning Research*, 2011, pp. 2825–2830.