音環境の類似度に基づいた会話場の認識と利用

中蔵聡哉[†]　角　康之[†]　西田豊明[†]

Development of the system detecting whom talking with by using sound field

Toshiya Nakakura,[†] Yasuyuki Sumi[†] and Toyoaki Nishida[†]

1. はじめに

人間のインタラクションを機械に理解させるという試みは盛んに研究されている. 中でもインタラクションの相手が誰であるのかというのは重要な情報であり, 赤外線や超音波等を用いて近くにいる人を検出しタグ付けするという試みは多数行われている. しかし, 位置関係が計測できたとしてもそれだけではどのような意味を持つのかは不明瞭である. 同じ距離にいても会話しているのと机に向かって作業を行っているのとではまったく異なる意味を持つからである.

我々はこのような物理量を利用して状況を推測する方法ではなく, 会話の場を直接意味的なまとまりに分類する手法を提案する. 意味的なまとまりの一例をあげると, 同じデモ発表を見ているというまとまりが考えられる. その場合発表者と聴衆全員が同じまとまりと考えるのが妥当であるが, 聞き手の一人が隣とおしゃべりをしていれば, さらにその二人をまとめるというように段階に分けて分類を行う.

このようにコミュニケーションの変化に柔軟に対応して分類を行うためには, 機械的に何かを発生させる手法では難しく, 人間がコミュニケーションで利用しているメディアを直接測定する必要がある. 今回我々は会話場検出の 1 段階目として, 会話をしていればお互いの声が聞こえているという当然の前提を利用し, 各ユーザの持つマイクの入力音声を比較することで同じ会話に参加しているかどうかの 2 値で判定するというシステムを構築し試行を行った.

ユーザの聞いている音に着目して関係を求めるシステムとして sociometer[1] があげられる. このシステムは長期間のユーザの行動の分析に着眼点があり, データを取りためた後の人手によるバッチ処理的分析を支援するものである. 我々のシステムは, 自動的にリアルタイムな判定を行い, 他のシステムの動作を支援する事を目的としている.

† 京都大学情報学研究科
Graduate School of Informatics, Kyoto University

2. システムの構成と動作

2.1 構　成

このシステムはモバイル端末上で動作し, 無線 LAN によるアドホックネットワーク上で P2P 通信を行い録音した音声を比較する. ユーザはモバイル端末を持つだけでよく, 自由に行動することができる. システム構成は図 1 の通りである.

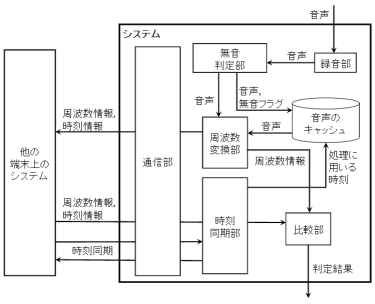

図 1　システム構成図

2.2 時刻同期

同じ会話であっても比較する時刻が異なれば音はまったく異なるので, マシン間の内蔵時計のズレの問題を解決しなければならない. 紙面の都合上詳細な説明は割愛するが, 簡易的な NTP を実装し 10ms 以下の誤差で時間差を計測する. この程度の差であればフー

リエ変換を行う際の窓関数で吸収される.

2.3 会話場判定アルゴリズム

基本的な考え方は, 入力音声が似ていれば同じと判断するという事である. また, 入力が無音であれば会話が行われていないのは自明であるため比較は行わない. 有音か無音かはマイク入力の音量を用いて判断を行う. 詳細なアルゴリズムは以下の通りである.

送信側の処理

1. 音声をバッファに蓄積しながら 3 秒待機する.
2. 1. で待った 3 秒の間に有音区間があれば 3. へ. なければ 1. へ戻る
3. 有音区間を最も長く含む連続区間 (1 秒間) の音声をフーリエ変換し, 時刻情報を付加してブロードキャストする. 送信後 1. へ戻る

受信側の処理

1. データ受信するまで待機する
2. 受信パケットのタイムスタンプを取得する.
3. スタンプと同時刻の受信側の音声が無音区間であれば 4. へ. 有音区間であれば 5. へ
4. 無音であれば話していない, すなわち異なると判断できる.1. へ
5. 受信周波数情報と自分の周波数情報を比較する
6. 類似度が閾値以上であれば同じ, 未満であれば異なると判定する.1. へ

比較に用いる周波数帯は人の声の周波数帯と言われる 100Hz〜4000Hz で, これを 3901 次元のベクトルととらえてコサイン類似度を求める.

3. 利 用 例

このシステムは, マイクを持ち, 通信可能な端末上であれば汎用的に利用できる. 現在 PhotoChat[2] というシステムで利用されている. PhotoChat は写真撮影と手書きメモを融合させ, 複数のユーザ間で共有することで, グループ内での各ユーザの興味への「気づき」の共有を加速し, その上での会話を促すことを目的とするツールである. このシステムの抱える問題点として, 写真が増えてきた際の整理の問題がある. 我々のセンサを利用することで, 「近くに誰がいたか」ではなく「誰と会話している状況で」撮った写真であるのか, というタグを付与することができ, おおまかな状況を把握して検索性を上げることができる.

4. 実 験

このシステムの評価実験を行った. 実験参加者はそれぞれがこのシステムをもち, プレゼンテーションを聞いた. またプレゼンの合間の休憩時間にはめいめいが個別に会話を行った. 実験中は各マシンでマイク入力を録音し, 同時刻のシステムの出力をログに残した. 人間が録音ファイルを聞いた時の判断を正解データとし, ログとの比較を行った.

その結果, 適合率はほぼ 100%, 再現率は 70%程度であった. 再現率が低くなっているのは主にプレゼン発表を聞く際の問題で, 聞こえてくる音が小さいと, 手元で作業をする音や咳払い, 録音ノイズ等が影響した瞬間は異なると判定されている. コミュニケーションの形態はそう早く変化するような類のものではないと考えられるので, 同じと判定されてからしばらくは同じであると丸める等の工夫が考えられる. 適合率が高いためこの丸めは正当である可能性が高く, また逆に考えればこれはプレゼンを集中して聞いていないというコンテキストの抽出であったとも考えられ, 丸めの程度によって会話場の段階分けが行える可能性がある.

個別の会話は, 近距離の会話であれば総じて正確な判定を行えているが, 約 3m 以上の距離での会話はノイズに埋もれ始め, 機械的に認識するのは難しくなる. 音量は距離の 2 乗に反比例して減衰してしまうが, 人間はそう感じていないため会話が行われない距離ではない. ただし,3m 以上の距離での会話は呼びかけ等短時間に終わることが多く, 長時間の会話になれば通常は近づく事になるので,1m 以内での会話と同程度に扱う必要はないと考える.

5. ま と め

意味的なまとまりをセンシングするシステムを開発した. このシステムはリアルタイムに出力を返し, 他のシステムに組み込んで利用することができる. 実験によって, このシステムはコミュニケーションの動的な変化に対応可能であることが確認された.

また今回のシステムは 2 値に分類するものであったが, 実験中に見られた現象によって, 同じと判定される頻度を用いれば会話場の強さの段階分けまで行える可能性があることが明らかになった. 今後はこの可能性を検証していく予定である.

参 考 文 献

1) Tanzeem Choudhury. Sensing and Modeling Human Networks. Doctoral thesis, Massachusetts Institute of Technology, September 2003.
2) 伊藤 惇, 角 康之, 久保田 秀和, 西田 豊明: 写真と書き込みの実時間共有による学会参加者間のコミュニケーション支援, 人工知能学会第 21 回全国大会, 2B4-1, 2007 年 6 月宮崎.