端末音声の相互相関に基づくアドホック会話の検出

Adhoc-Meeting Detection Using Cross-Correlation of Terminal Audio Data

岡本 昌之 ♥ 池谷 直紀 ♦
西村 圭亮 ♠ 菊池 匡晃 ♣
長 健太 † 服部 正典 ††
坪井 創吾 ††† 芦川 平 ††††

Masayuki OKAMOTO Naoki IKETANI
Keisuke NISHIMURA Masaaki KIKUCHI
Kenta CHO Masanori HATTORI
Sougo TSUBOI Taira ASHIKAWA

複数の個人携帯端末で記録した音声情報をサーバに集約し，音声の特徴量データ同士の相互相関計算に基づきユーザ間の会話状況を検出する手法を提案する．本手法は，生の音声データをアップロードすることなく各ユーザの会話の有無を推定することが特徴である．オフィス環境における，8 人のユーザの実運用を通じて収集した合計 4,605 個の正解データを用いた評価の結果，F 値の平均 0.9 の精度で会話状況を検出可能と分かった．

We propose a method to detect adhoc meeting based on cross-correlation between audio feature data, which are collected from personal mobile terminals. This method can detect whether there is conversation between each pair of users without raw audio data. Through a two-day evaluation with 8 users, we found our method could detect meeting contexts with 0.9 F-measures on average.

1. はじめに

本稿の目的は，コミュニティ分析支援や会話状況を属性として用いるアプリケーションのための,「いつ」「誰と誰が」会話したかを検出する簡便な方式の提案である．

♥ （株）東芝 masayuki4.okamoto@toshiba.co.jp
♦ （株）東芝 naoki.iketani@toshiba.co.jp
♠ （株）東芝 keisuke1.nishimura@toshiba.co.jp
♣ （株）東芝 masaaki11.kikuchi@toshiba.co.jp
† （株）東芝 kenta.cho@toshiba.co.jp
†† （株）東芝 masanori.hattori@toshiba.co.jp
††† （株）東芝 sougo.tsuboi@toshiba.co.jp
†††† （株）東芝 taira1.ashikawa@toshiba.co.jp

これまで，人々の様々な会話状況を記録・分析する研究が多数行われてきた．会話状況記録の目的は主に議事録・発言録の作成自動化や発言の分析に関するものと，会話が行われた事実の記録に関するものとに大別できる．我々のアプローチは主に後者を対象とする．

前者は，例えば会議記録作成に関してはアノテーションベースの手法 [1] や音声・映像認識による手法 [2] を用いた研究がある．これらの取り組みは基本的に会議室に設置された機器を用いて会議自体の内容を詳細に収録・分析するものである．後者は，日常生活で誰と誰が会話したかを記録・可視化することを主な目的とし，例えば人間関係や会話の分析に用いられる．会話の検出は，前者と同様に会議参加者を記録・蓄積する方法や，IC，赤外線タグを用いた会話相手の検出方法 [3] が用いられる．

しかしながら，後者の会話記録を日常的に行う場合，以下の課題がある．まず，会話は会議室だけで行われるわけではなく，会議外を含めた会話も記録することが必要である．例えば，定例の進捗報告会議では参加者全員での情報共有が行われるが，会議外でも，より詳細な内容に関しては各分担の担当者同士の少人数の会話が行われる．このような会話を本稿では**アドホック会話**と呼ぶ．アドホック会話を含む様々な会話状況を記録するためには，会議室に特別な機器を設置するのではなく，個人に紐付いた機器を用いて記録することが必要となる．

RFID タグや赤外線タグを用いた手法 [3] はより個人に紐付いた手段であるが，RFID タグではユーザ同士が近い場所にいるかどうかは判別できるものの，会話したことは検出できない．また，赤外線タグを用いた対面状況の検出手段では，1 対 1 など少人数の会話は検出しやすいが多人数の会話は検出が困難となる．

このような課題を解決するために，我々は個人がそれぞれ所有する端末で記録される音声情報をサーバに集約し，音声の特徴量データ同士の相互相関計算に基づき会話状況の検出する手法を提案する．提案方式の特徴は，生の音声データをアップロードすることなく各ユーザの会話状況の有無を推定すること，および各端末の時刻が完全に同期していなくても検出できることである．

また，本方式は単なる分析ツールへの適用だけでなく,「いつ」「誰が」会話したかを属性として用いることにより既存の会議支援システムやグループウェアに応用可能である．例えば，最近ではノート PC を用いた手持ち資料の提示・説明は広く行われるが，後から「この資料は誰に見せたのだろうか？」「先日○○さんに見せた資料はどこにある？」のように，会話を手掛かりとした資料検索が必要な場面も増えている．提案方式の併用により，デスクトップ検索のクエリとして人名が入力可能となる．

次節以降，提案手法の詳細を述べるとともに，ノート PC を用いた実運用に基づく評価結果を報告する．また，会話状況検出の応用例として，デスクトップ検索への適用について紹介する．

2. 会話状況検出方式

本節では，我々が提案する会話状況検出システムの構成と検出方式を述べる．

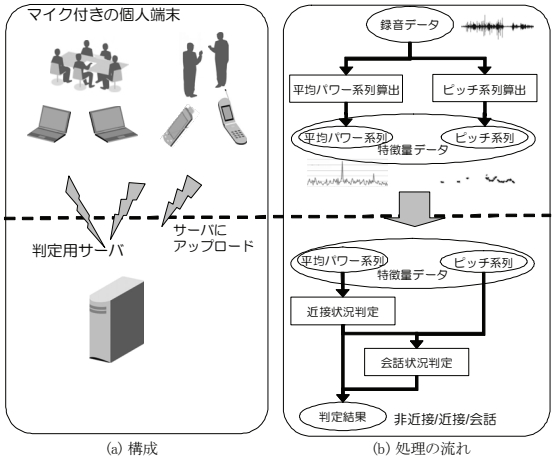


図 1　会話状況検出システム
Fig. 1　Meeting context detection system.

本システムは，図 1(a) のようにマイクを備えた個人端末と判定用のサーバから構成される．このシステムを通じて，図 1(b) に示すように録音データから会話に関連する状況が検出される．

本稿では，検出対象の状況として，次の 2 つを導入する．ユーザ同士が同じ環境音を共有する近隣にいる場合を**近接状況**にある，と定義する．「近隣」を示す距離としては，Hall の定義する社会距離から公衆距離の近接相までを想定する [4]．ユーザ同士が近接しており，かつ会話している場合を**会話状況**にある，と定義する．「会話」にはすれ違いざまの挨拶のようなものは含まず，話題を構成する程度の長さ（10 数秒以上）を想定する．なお，本稿では電話での会話は対象外とする．

2.1　処理の流れ

本節では，図 1(b) とあわせて会話状況検出の流れを説明する．

まず，ユーザは各自が所持するノート PC や携帯電話，IC レコーダなどのマイク付き機器により常時録音し，機器上のクライアントは特徴量である**平均パワー系列**と**ピッチ系列**をタイムスタンプと合わせて記録する．クライアントは定期的に特徴量データをサーバにアップロードする．サーバは，特徴量データがアップロードされると，既にアップロードされた特徴量データのうち，近いタイムスタンプを持つ他の 1 ユーザのデータを選択し，パワー系列を用いて近接状況にあったかを判定する．近接状況にあると判定された場合，さらにピッチ系列も用いて会話状況にあったかを判定する．最終的に，非近接/近接/会話のいずれかの状況が判定結果として出力される．

音声を各端末からサーバに集約する際，本システムでは録音データそのものではなく，特徴量データのみをアップロードする．本方式は録音データそのものをアップロードする場合と比べ，以下の特長がある．

少ないネットワーク・CPU 負荷：音質をある程度落としても，電話程度の音質では数 10kbps の帯域が必要だが，本方式では特徴量のみ利用するためより少ない帯域で済む．本稿での実装とパラメタではクライアントあたり 2kbps 程度で特徴量データが送信される．また，サーバ側の判定処理コストも録音データそのものを処理する場合と比べ小さい．

プライバシー侵害の防止：常時高速回線が利用できるなど，帯域が十分な環境下であっても，録音データ自体をアップロードする場合ユーザのプライバシーを侵害する可能性があるが，本方式ではサーバで発話内容自体を知られる心配はない．

2.2　検出方式

本節では，クライアント側で行う特徴量算出，およびサーバ側で行う近接・会話状況検出の手順を説明する．

クライアント側では以下の特徴量算出が行われる．

まず，予め決められたシフト，フレームで，音声の平均パワーを計算し，元の音声から平均パワーの系列を得る．時刻 t に入力される音量を $l(t)$ とすると，時刻 t_i から始まる区間に対して，フレーム f の幅から得られる平均パワー $P(t_i)$ は $P(t_i) = \sum_{t=t_i}^{t_i+f-1} l^2(t)/f$ となる．

次に，ピッチ推定を行う．推定手法としては規格化相互相関 (normalized cross correlation)[5] を用いた既存の手法 [6] によりピッチが適当な周波数領域に含まれる区間を抽出し，ピッチ系列を出力する[1]．

上記処理により得られた平均パワー系列，およびピッチ系列がサーバにアップロードされる．サーバでは以下の処理が行われる．

まず，2 つの平均パワー系列の相互相関に基づく確信度を元に近接状況判定を行う．処理手順を以下に示す．

(1) 2 人のユーザ u_1，u_2 にそれぞれ対応するパワー系列 w_1, w_2 について，w_1 の時刻 t_0 からの N 単位に対し w_2 の時間差 m の N 単位と規格化相互相関 $R_{fg}(t_0, m)$ を計算する．

(2) m を $-M \leq m \leq M$ の範囲でステップ t_{step} で変えながら繰り返し，最も相互相関が高くなる m を見つける．$m > 0$ は w_1 に対して w_2 が遅れていることを，$m < 0$ は w_2 に対して w_1 が遅れていることを表す．相互相関を最も高くする m が $w1$ の時刻 t_0 に対する $w2$ の時間差とみなせる．

(3) 時刻 t_0 を変えながら (1)，(2) を繰り返すことで，時間差の分布を示すヒストグラム $d(t)$ $(-M \leq t \leq M)$ が得られる．

(4) ピーク値となる時間差を与える比率 $r_{conf} = \max(d(t))/\sum d(t)$ が繰り返しのうち一定の比率 r_{th} を超え，かつ相互相関の平均 R_{av} が閾値 R_{th} 以上の場合に近接状況とみなす．以降ではこの比率 r_{conf} を確信度と呼ぶ．

なお，端末毎の時刻の時間差およびネットワークの遅延は，この計算の過程で計算される時間差の範囲 $[-M, M]$ が十分大きければ確信度計算の段階で吸収される．

上記処理で近接状況と判定された場合，さらにピッチ系列から音声らしい区間の長さ，および平均パワーの大きさが閾値を超えたか否かにより会話状況判定を行う．ピッチ推定の結果は，入力された音の**音声らしさ**を表し，平均パワーは**端末所有者の発話ら**

[1] 本稿では，「ピッチ」を有声音の基本周波数 (F0) の別称として用い，聴覚分野における心理量としてのピッチとは区別する．

しさを表す．両者を併用することで会話らしさを検出する．音声らしさは与えられた周波数の範囲に含まれるピッチの区間の長さ（本稿ではこの区間を**発話候補区間**と呼ぶ）で判定する．両者が満たされた場合に会話状況と判定する．

手順 (4) では，手順 (3) により得られる時間差の分布において，ピークとなる時間差のみ調べる場合と，周辺まで見る場合とが考えられるが，本稿では，隣りあった時間差の値まで含めて数える．これは，ノイズの影響や 2 つのマイクが接続された PC 自体の個体差による時間のずれを吸収するための方策である．

判定結果は，2 人の組に対する単位時間あたりの結果として出力される．以下，近接状況・会話状況に関する判定結果のデータをそれぞれ**近接判定データ**および**会話判定データ**と呼ぶ．

2.3 実装

会話状況判定アプリケーションのクライアント，サーバを以下の通り実装した．

クライアントは Windows XP 上の常駐アプリケーションとして動作する．前章の方式に加え，クライアント間の時間差を最小限にするための NTP 機能，およびネットワークが利用できない場合にアップロード用の特徴量データをキューに保存し，ネットワーク接続時に順次アップロードする機能を持つ．

サーバはクライアントから特徴量データがアップロードされると，特徴量データに含まれるタイムスタンプを確認し，同時間帯の他ユーザの特徴量データと判定を行う．近接・会話状況の検出結果はサーバに記録され，他の端末から自分のユーザ ID と時刻を入力すると会話相手のユーザ ID を返す．

3. 評価

本節では，2 節で述べた方式がどの程度有効にはたらくか調べるために，主にオフィス環境で活動する 8 人のユーザが実際に 2 日間運用し，評価した．

3.1 評価手順と実験設定

以下の手順により評価を実施した．

まず，各ユーザは，所有するノート PC 上で対象期間中会話状況検出クライアントを常駐起動させる．ユーザは適宜 Web ブラウザを用いて近接・会話状況の判定結果の閲覧・確認用サーバにアクセスし，正解データの入力を行う．近接・会話の各状況に関しては，ユーザのノート PC の配置を基準に入力する．入力は各ユーザ毎のペアに対して 1 分単位で行う．つまり，8 人全員がクライアントを起動中の場合は，あるユーザについて，同じ 1 分間に 7 組のペアに対応する判定結果 7 個が存在する．近接状況とみなす基準に関しては，オフィスにおいて普段作業する範囲を元に決定した．オフィスにおける座席配置は最も遠いユーザ間で 6m 程度あるため，この範囲を基準とした．また，会話状況とみなす基準に関しては，会話参加者のノート PC が各自の手元にある状況で会話した場合のみ，会話状況であるとみなした．

評価データの収集後，平均パワーの閾値，および発話候補区間の閾値を変化させつつ，それぞれの近接判定データ，会話判定データと，ユーザが入力した正解データとを比較することで提案方式の精度を調べた．指標として再現率，適合率，F 値を用いた．

また，実験を行った環境およびパラメタは次の通りである．

実験は主にオフィス内で行われ，主に居室での個人作業や 2〜3 名でのアドホック会話，会議室での会議（ユーザ以外が参加する会議も含む）が行われた．また，数名のユーザについては外出先での記録データも一部含まれる．

クライアントとして，各ユーザの所有するノート PC を用いた．録音は全て PC の内蔵マイクを用い，処理前の音質は周波数が 8kHz，量子化ビット数が 16bit のモノラルとした．また，特徴量データ算出のパラメタとして，パワー系列計算時のフレーム幅を 0.125 秒，シフト幅を 0.05 秒とした．サーバにおける近接状況判定の計算におけるパラメタとして，$N = 40$，$M = 60$（秒），$t_{step} = 0.1$（秒）を用いた．また，ピッチ推定のパラメタとして，F0 の検出範囲を 50〜550Hz，フレーム幅を 0.0075 秒とした．判定を行う区間を 1 分単位としているため，$t_{step} = 0.1$ より確信度計算の分母は 600 となる．クライアント・サーバともパラメタは予備調査の結果，決定した．

サーバ側での近接・会話状況判定処理にかかる時間は，Pentium 4 3.80GHz Dual の CPU を搭載した PC の Windows XP Professional Service Pack2 上で動作させた場合で特徴量データ 1 組あたり平均約 0.4 秒である．

評価用データの収集作業を 2007 年 10 月 24 日，25 日の 2 日間実施し，正解付きの近接判定データを 6,397 個，会話判定データを 4,605 個収集した．会話判定データに関しては，正解入力に確信を持てず未入力のものが含まれるため，近接判定データよりも正解データが少ない．

3.2 結果

本節では，近接状況，会話状況に関するそれぞれの結果を示す．

3.2.1 近接状況検出

まず，近接状況の検出結果を示す．今回の正解データでは，近接状況である場合が全体の 8 割以上を占めたため，近接状況の F 値をそのまま調べることの意味は小さい．したがって，近接状況と非近接状況それぞれの再現率を調査した．図 2 に確信度の閾値と再現率の関係を示す．

図 2 より，どのユーザについても確信度 20 を境界とする急な勾配があり，この値を閾値とする時に近接状況・非近接状況のバランスが良いことが分かる．再現率，適合率，F 値を表 1 に示す．

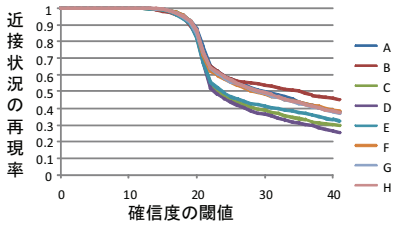
表 1 より，近接状況である割合が多いことを差し引いても高い適合率・F 値を達成することが分かる．

3.2.2 会話状況検出

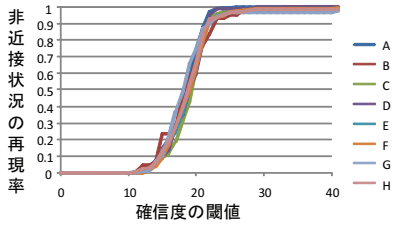
次に，会話状況検出の結果を示す．まず，再現率・適合率・F 値の比較を表 2 に示す．これらの値は，それぞれのユーザについて F 値が最も高くなるパラメタを用いた場合の結果である．

表 2 より，F 値で 0.77 以上，特に 8 人の平均では約 0.9 と高い精度で会話状況を検出することが分かる．

次に，この結果において，パワーによる閾値の効果と発話区間候補比率の閾値による効果のいずれが寄与するかを調べた．典型的な結果のパターンを代表してユーザ E，G，H の比較を図 3 に示す．横軸は 1 分間の平均パワーの常用対数に関する閾値を示



(a) 近接状況の再現率



(b) 非近接状況の再現率

図 2　近接状況判定の結果

Fig. 2　Result of proximity context detection.

表 1　近接状況の再現率，適合率，F 値（確信度の閾値 20）

Table 1　Recall, precision, and F-measure of proximity context (threshold of confidence score is 20).

ユーザ	データ数	近接状況の数	再現率	適合率	F 値
A	1710	1514	0.87	0.96	0.92
B	494	451	0.84	0.96	0.90
C	1839	1695	0.85	0.96	0.90
D	1316	1214	0.83	0.96	0.89
E	1715	1380	0.82	0.91	0.86
F	1589	1343	0.86	0.93	0.89
G	1514	1427	0.85	0.98	0.91
H	2047	1868	0.87	0.96	0.91

し，縦軸は F 値を示す．青線，赤線，緑線は，それぞれ平均パワーの対数のみを用いた場合，ピッチのみ用いた場合，両方組み合わせた場合をそれぞれ表す．事前評価として，各ユーザにつき発話候補区間の閾値を 50, 100, 150 （区間全体は 600）と変化させた結果，発話区間候補の閾値を 100 に設定した場合に良好な結果が得られたため，この数値を用いた．

図 3 より，以下のことが分かる．

まず，発話候補区間による閾値変化の F 値への影響はパワーによる閾値設定と比べると緩やかなため，少なくとも本実験のように，利用環境がある程度共通している場合は全員共通のパラメタを利用可能である．

また，全体的に，会話状況検出への貢献度は平均パワーよりもピッチによる割合が大きいが，他人の声が入るなどの場合を考慮すると，パワーを利用することでピッチのみの場合よりも高い効果が得られる場合がある．特に，ユーザ G，H においてはその効果が大きいことが分かる．ただし，閾値を過度に高く設定すると再現率が下がる可能性がある．

表 2　会話状況検出結果の比較（各自の F 値が最も高いパラメタを利用．発話区間候補の閾値: 100/600）

Table 2　Comparison between meeting-detection results (using parameters with the maximum F-measure. The threshold of pitch occurrence rate: 100/600).

ユーザ	データ数	会話状況数	再現率	適合率	F 値
A	1014	179	0.93	0.87	0.9
B	343	139	0.9	1.0	0.95
C	905	224	0.92	0.92	0.92
D	913	141	0.89	0.97	0.93
E	1261	244	0.89	0.96	0.92
F	1475	166	0.93	0.94	0.93
G	1174	124	0.94	1	0.97
H	1775	145	0.72	0.83	0.77

3.3　考察

今後，提案手法を広く適用する場合には以下の事項を検討する必要がある．

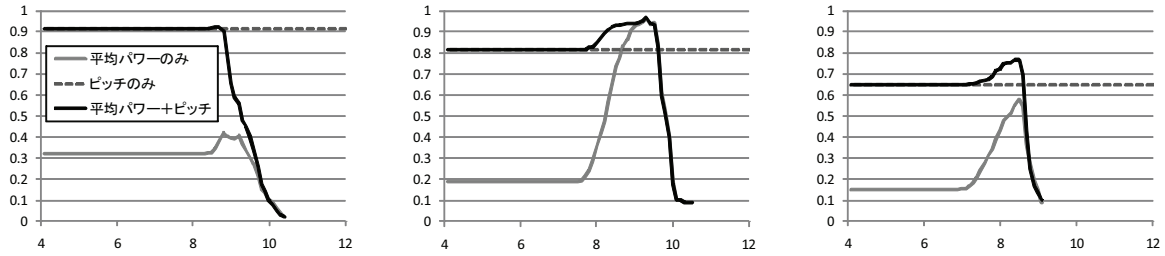
環境変化とパラメタ調整について

提案手法では，近接判定における確信度の閾値，発話候補区間の閾値，平均パワーの閾値と，3 種類の閾値を設定する必要がある．評価の結果，同じ環境を想定する場合には，確信度および発話候補区間の閾値に関してはユーザ間で共通のパラメタを利用可能と考えられる．平均パワーに関しては，マイクのゲイン設定の違いに加え，ユーザ毎の発話音量に個人差があるため調整が必要と考えられる．オフィス環境を想定する場合には，閾値が大きく変動するとは考えにくいため，一度決定すれば問題ないと思われる．ただし，極端に背景ノイズの音量が変わる移動を伴う場合は，パワー分布を元に閾値を調整する，ピッチによる推定結果のみ利用する方式に切り替える，などの手法が必要と考えられる．

対応するユーザ数について

提案手法では 2 人の組に対して計算を行うため，対象となるユーザ数が増えると組み合わせにより計算量が増大する．2.2 節より，サーバ側での判定コストは近接状況判定の相互相関計算がほとんどを占めるが，この部分の計算コストは ${}_nC_2 \times M \times N \times 1/t_{step}$ に比例する（n はユーザ数）．ここで，3.1 節より，$M = 60$ 秒の場合に 2 人の 60 秒分のデータの組を判定するのにかかる時間が Pentium 4 3.80GHz Dual 程度のスペックの PC を用いた場合約 0.4 秒であるが，NTP を用いて誤差 1 秒以内に時刻が同期された PC を対象とする場合は M をより小さな値（例えば $M = 5$ 秒）としても構わない．会話状況判定を含むコストやオーバヘッドを考慮してこの処理が約 1/10 と想定すると，例えば 55 人分の 60 秒のデータを処理する時間は $0.4 \times {}_{55}C_2 \times 1/10 = 59.4$ 秒程度となり，この人数であればほぼ遅延なくデータを処理可能と分かる．

より多数のユーザを扱う場合，例えば判定結果の連鎖的な適用



(a) ユーザ E，閾値 100　(b) ユーザ G，閾値 100　(c) ユーザ H，閾値 100

図 3　F 値の比較例（ユーザ E，G，H）．発話候補区間の閾値は全体を 600 としたもの．横軸は平均パワー閾値の常用対数，縦軸は F 値を表す
Fig. 3　Comparison of F-measure (users E, G, and H). The horizontal axis is common logarithm of the threshold of average power and the vertical axis is F-measure.

による計算コスト削減が考えられる．例えば，ユーザ A とユーザ B，ユーザ B とユーザ C がそれぞれ会話状況であれば，ユーザ A とユーザ C も会話状況であるとみなせるため，計算を省略可能である．あるいは，サブネットが物理的な位置関係と対応する環境であれば，リアルタイムに送信されるデータに関しては，サブネット間で近接/会話状況検出を行う必要はない．

環境音が共通する場合について

提案手法では，同じテレビ番組を視聴するなど同じ環境音を共有する場合，実際には離れていても近接状況であると判定される可能性がある．近接状況でないことを音声だけから明確に知ることは簡単ではないが，判定前後の特徴量の比較や，ネットワークに接続されている場合はサブネットの違いの検出によりある程度誤検出を防ぐことが可能と考えられる．

ただし，同じ環境音を共有することは，同じ状況を共有することでもあると考えられるため，用途によっては有用な情報と言える．

4.　応用：人をクエリとしたデスクトップ検索

本稿で述べた会話状況検出方式は主に組織内での活動記録での分析用途での利用が想定されるが，我々は利用者をより積極的に支援する機能としても用いたいと考えている．本節では，検出された会話状況を時間と人の組として表される属性ととらえた場合の応用例として，ファイルの活用履歴に注目したファイル検索アプリケーションとの連携について述べる．

3 章で報告した会話状況検出結果とファイル操作履歴とを対応付けることにより，あるファイルを閲覧した時に誰と会話していたかを知ることができる．我々は，ファイルの作成・コピー・削除，Microsoft® Word/Excel/PowerPoint 形式のファイルに対する各操作（例えば閲覧・スライドショー表示）等，ファイル操作が時刻と対応付けて記録するファイル操作履歴記録ツールを実装し，さらに操作履歴とファイル操作時の会話相手を対応付けることで，誰にどのファイルを見せたかを検索するデスクトップ検索アプリケーションを実装した（図 4）．

本アプリケーションは次のように利用される．まず，ユーザは対象となる会話相手と，会話した時刻を含む期間を指定する．検索開始ボタンを押すと，アプリケーションは会話状況検出サーバに対し，期間中の会話相手と会話時刻の組を問い合わせる．会話時刻とファイル閲覧記録を照合し，対応付けることができたファイルの一覧が結果として出力される．また，検索条件としてユーザ ID の否定表現を用いることで，ある一定期間に相手にまだ見せていない資料の一覧を知ることも可能である．

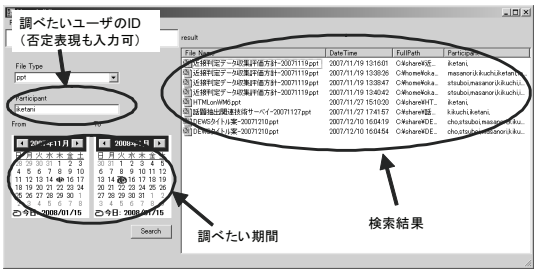


図 4　ファイル検索アプリケーション
Fig. 4　Snapshot of a file search application.

5.　関連研究

本稿で述べた相互相関を用いた近接判定処理は，ノイズを含む類似音声のマッチング処理であると捉えることもできる．音声処理分野では，楽曲検索等でこのような取り組みがなされている [7]．最近では画像処理ベースの手法を用いた高速なマッチング手法も開発されており [8]，視聴中のテレビ番組とのマッチングに関する応用 [9] にも利用されている．これらの応用は，正解となる楽曲や番組音声のデータに対しどれだけノイズや歪みを含んだ音声で検索できるかというものであり，本稿の提案方式とは用途が異なる．しかしながら，今後精度向上などを考慮する場合にはこれら既存手法との比較も必要と考えられる．

また，会議を含む会話情報の記録は様々な方式が提案されている．会議記録作成に関しては，アノテーションにより発言の手掛かりを記録する方式 [1] や，音声・映像認識による記録方式 [2] 等が提案されている．また，対面状況を利用・蓄積する研究として，赤外線タグを用いた手法 [3] や，ウェアラブル機器と組み合

わせたインタラクションコーパス収集 [10] などがある．これらは主にイベント会場など特定の場所での利用を対象とするが，ビジネス顕微鏡 [11] のように，日常活動において対面状況を蓄積する試みもある．本稿の手法と比べ，これらの手法はまず赤外線を用いて対面状況を検出する必要があるが，デバイスやインフラが揃っている環境であれば比較的簡単に会話相手を識別できるため，補完的な利用も考えられる．

6. おわりに

本稿では，個人携帯端末で記録した音声特徴量をサーバに集約し，相互相関計算に基づきアドホックに発生するユーザ間の会話状況を検出する手法を提案した．本手法は，生の音声データをアップロードすることなく各ユーザの会話状況の有無を推定することが特徴である．ノート PC を用いた 8 人による 2 日間の実運用に基づく評価結果を通じ，平均して 0.9 以上の高い F 値を達成する精度で会話状況を検出可能と分かった．また，検出された会話状況をファイルへのアノテーションに用いる文書検索への応用事例を紹介した．

今後，より多くのユーザによる大規模評価を行うとともに，ファイル検索支援等のアプリケーション実装，グループ内・グループ間でのコミュニケーション分析も進める予定である．

[文献]

[1] K. Nagao, K. Kaji, D. Yamamoto, and H. Tomobe, “Discussion Mining: Annotation-Based Knowledge Discovery from Real World Activities,” in *Proc. Pacific-Rim Conf. on Multimedia*, 2004.

[2] Steve Renals and Thomas Hain and Hervé Bourlard, “Recognition and Understanding of Meetings The AMI and AMIDA Projects,” in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2007.

[3] R. Borovoy, F. Martin, S. Vemuri, M. Resnick, B. Silverman, and C. Hancock, “Meme Tags and Community Mirrors: Moving from Conferences to Collaboration,” in *Proc. ACM Conf. on Computer Supported Cooperative Work*, pp. 159–168, 1998.

[4] E. T. Hall, *Hidden Dimension*, Doubleday, 1966.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons Inc., 1973.

[6] D. Talkin, “A Robust Algorithm for Pitch Tracking,” *Speech Coding and Synthesis*, pp. 495–518, 1995.

[7] J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System,” in *Proc. Int. Conf. on Music Information Retrieval*, 2002.

[8] Y. Ke, D. Hoiem, and R. Sukthankar, “Computer Vision for Music Identification,” in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2005.

[9] M. Fink, M. Covell, and S. Baluja, “Social- and Interactive-Television Applications Based on Real-Time Ambient-Audio Identification,” in *Proc. European Conf. on Interactive TV*, 2006.

[10] Y. Sumi, S. Ito, T. Matsuguchi, S. Fels, and K. Mase, “Collaborative capturing and interpretation of interactions,” in *Proc. Pervasive 2004 Workshop on Memory and Sharing of Experiences*, pp. 1–7, 2004.

[11] 矢野 和男, 栗山 裕之, “「人間×センサ」センサ情報が変える人・組織・社会,” 日立評論, Vol. 89, No. 07, pp. 572–577, 2007.

岡本 昌之 Masayuki OKAMOTO
（株）東芝 研究開発センター知識メディアラボラトリー研究主務．2003 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主にコンテクストアウェア技術および情報抽出の研究開発に従事．情報処理学会，人工知能学会，ACM 各会員．

池谷 直紀 Naoki IKETANI
（株）東芝 研究開発センター知識メディアラボラトリー研究主務．1998 年慶應義塾大学大学院理工学研究科修士課程修了．エージェント技術，コンテクストアウェア技術の研究開発に従事．情報処理学会会員．

西村 圭亮 Keisuke NISHIMURA
（株）東芝 研究開発センター知識メディアラボラトリー所属．2005 年東京電機大学情報環境工学科卒業．コンテクストアウェア技術および知的生産活動における支援技術の研究開発に従事．情報処理学会会員．

菊池 匡晃 Masaaki KIKUCHI
（株）東芝 研究開発センター知識メディアラボラトリー所属．2006 年大阪大学大学院工学研究科修士課程修了．主にコンテクストアウェア技術および情報抽出の研究開発に従事．

長 健太 Kenta CHO
（株）東芝 研究開発センター知識メディアラボラトリー研究主務．1997 年早稲田大学大学院理工学研究科修士課程修了．エージェント技術，コンテクストアウェア技術の研究開発に従事．情報処理学会，日本ソフトウェア科学会，電気学会各会員．

服部 正典 Masanori HATTORI
（株）東芝 研究開発センター知識メディアラボラトリー研究主務．1996 年九州大学大学院工学研究科修了．博士（情報科学）．2005 年〜2006 年米国マサチューセツ工科大学客員研究員．エージェント技術，コンテクストアウェア技術などの研究開発に従事．情報処理学会，ACM 各会員．

坪井 創吾 Sougo TSUBOI
（株）東芝 研究開発センター知識メディアラボラトリー研究主務．1999 年東京工業大学大学院総合理工学研究科修了．CSCW，コミュニティ支援に関する研究開発に従事．人工知能学会会員．

芦川 平 Taira ASHIKAWA
（株）東芝 研究開発センター知識メディアラボラトリー所属．2003 年九州大学大学院システム情報科学府修了．主に情報共有システムの研究開発に従事．