身体的拘束に基づく音声駆動体幹動作生成システム

Speech Driven Trunk Motion Generating System Based on Physical Constraint

○境　くりま [*1,2], 港　隆史 [*1], 石井　カルロス寿憲 [*1], 石黒　浩 [*1,2]
Kurima SAKAI[*1,2], Takashi MINATO[*1], Carlos Toshinori ISHI[*1], Hiroshi ISHIGURO[*1,2]
ATR[*1], 大阪大学大学院 基礎工学研究科 [*2]
sakai.kurima@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp, minato@atr.jp, ishiguro@sys.es.osaka-u.ac.jp

Abstract

近年，様々なヒューマノイドロボットが開発されてきており，人の代わりとなり社会的な役割を果たすことが期待されている．ヒューマノイドロボットが人間らしい動きをすることで，我々はロボットに対し親密感を覚える．特に，人々に受け入れられる対話ロボットを実現するためには，発話に伴う動作が必要となる．本論文では，人と対話するヒューマノイドロボットの頭部，腰部動作に着目し，ヒューマノイドロボットの発話に合わせて，人らしい頭部，腰部動作をリアルタイムで生成するシステムを構築する．

1　はじめに

近年通信技術やセンサ技術の発達によりロボットがより身近なものになってきた．特にヒューマノイドロボットは，遠隔操作することで場の共有感や身体動作といった非言語情報を伝達することができるため，電話やビデオチャット以上に遠隔地の人と直接対面しているような対話が実現できる [1]．特に，人手不足が深刻な高齢者介護の現場では，高齢者と遠隔地の人をつなぐことで役立っている [2]．また，自律ヒューマノイドロボットによるイベント会場の案内役 [3]，デパートでの販売員 [4]，病院での陪席者 [5]，や受付 [6] など社会的役割を人の代わりに果たそうという試みも行われている．以上のようにヒューマノイドロボットには，人の代わりとなり社会的な役割を果たすことが期待される．

ここで問題となるのは，人々に受け入れられるためのロボットの振る舞いのデザインである．人はエージェントの外見からその振る舞いを予測し，人間らしい見た目には人間らしい振る舞いを期待する傾向にある（適応ギャップ）[7]．特に，人間に外見が酷似したアンドロイド（図 1）に対して，それに応じた人間らしい動きを期待する．アンドロイドが期待される振る舞いを行わなければ，悪い印象を与えることとなる．実世界で動くアンドロイドでは，アクチュエータの自由度などのハードウェア的な制約があり，人間と同一の動きが実現できないため，人がどのような動きに人間らしさを感じるのか，その要素を明らかにして動きをデザインする必要がある．また，人間らしい動きは，外見にかかわらず人型エージェントに対する親密度を向上させることが報告されており [8]，人間らしい動きを感じさせる要因を明らかにすることは，人型エージェント全般において意義がある．

図 1: Android ERICA

従って，人々に受け入れられる対話ロボットを実現するためには，発話に伴ってどのような動作を表出すべきかが課題となる．対話ロボットにおいて，人らしさの要因として最も重要な点は，ロボット自身が発話しているという印象である．その印象を与えるための基本的な動作は，発声のための運動である．発声のための動き（口唇動作だけでなく，首，胸，腹の動き）が，発声と同期して表出されれば，ロボット自身が発話しているという印象を強める．人の発話と動きの関係をモデル化し，発話情報から動作を自動生成すれば，最も基本的な発話時の人らしい振る舞いとなる．本研究では，人と対話するヒューマノイドロボットの頭部，腰部動作に着目し，ヒューマノイドロボッ

トの発話に合わせて，人らしい頭部，腰部動作をリアルタイムで生成するシステムを構築する．

2 関連研究

コンピューターグラフィックスの研究分野では，エージェントの発話に合わせ頭部動作を自動生成する手法がいくつか提案されている．Le et.al. は発話音声のパワー，ピッチと頭部の 3 自由度の動きを Gaussian Mixture Model を用いてモデル化し，リアルタイムで頭部動作を生成するシステムを提案している [9]．また，隠れマルコフモデルを用いた同様のモデル化も行われている [10, 11, 12]．しかし機械学習を用いた自動生成システムでは，学習に使われているモーションデータが収録された状況に合った動作しか生成できない．特に，対話相手との関係性により話し方が変化するため，すべての状況での動作を収録することは困難である．また，これら手法は収録されたデータを復元することを目的にしているため，異なる状況で使用するための動きの変調や他の動きと複合することができない．エージェントの動作は対話状況に応じて複数の動作をミキシングすることが重要になり，様々なミキシングの手法が提案されている [13, 14, 15]．そのため，エージェントの発話する動作のみに着目したシステムが必要となる．

本論文では日本語の発話に合わせた動作生成を扱うのに対し，上記の研究は主に英語を母国語とする動作生成手法である．日本語に対する動作生成もいくつか提案されている．Watanabe et.al. は，発話の on/off 情報から頷きのタイミングを推定する手法を提案している [16]．しかし，頷き生成のタイミングを生成するだけで，どのような関節の動きが人間らしさを生むかまでわかっておらず，実際のアンドロイドで使用するには不十分である．Ishi et.al. は，発話の意味に対する動作のマッピング方法を提案している [17, 18]．発話の意味を推定するためには，韻律特徴のみならず言語特徴も利用する必要があるため [19]，リアルタイムシステムを構築することが困難である．

一方で，解剖学の知見から，口の開閉動作に伴い頭部が動くことも報告されている [20]．この知見から頭部の発話動作も社会的状況の要素以外の身体的拘束をもとに生成できる可能性がある．

本論文では，社会的状況に依存せず，純粋に発話のための動作を，人間の身体的拘束を利用し発話情報に基づいてリアルタイムで生成することを目的とする．また，機械学習で構築したモデルでは，発話と動作のどのような特徴が人間らしさに関わっているのか，解析するのは容易ではない．本研究では，動作の要因が直感的に分かりやすい動作生成モデルの構築を目指す．特に，目線をそらす動作は対話のコンテキストに依存し [21]，そのパターンは個性に依存する [22] ことから，本論文では発話に合わせた首と腰の縦方向の動きに着目する．

3 韻律と頭部動作の関係見つける実験

本節では人間らしい発話動作を自動生成するためのルールを見つけるための実験を説明する．人間が発声する際頭部動作などが音声に同期することが報告されており，特にパワーとピッチの変化と動作の変化が同期することが知られている [23]．しかし，日本語ではパワー，ピッチの韻律特徴と頭部動作の相関は高くないことも報告されている [24]．また，解剖学の知見から，口の開閉動作に伴い頭部が動くことも報告されている [20]．そのため，従来の音声のパワー・ピッチに加え，口の開き度合の 3 要素が社会的なインタラクションを含まない状況でも動きと相関があるのかを明らかにする．

3.1 実験設定

口の開閉が母音を発音する際に大きく変化するため，実験参加者に「あ・い・う・え・お」を 3 秒間発声してもらい，その発声に伴う首の動きの変化を計測する．母音の発声はそれぞれを高音・中音・低音で発音する条件 (Voice Pitch Condition) と，発声しやすい声の高さで大声で発声する条件 (Mouth Openness Condition) を設けた．被験者には，各発声ごとに正面を一旦向くよう指示を出し，姿勢をリセットした．予備実験より，被験者は母音を発声する際に 2 要因（高音で大きな声など）を混同させると発声しづらかったため，本実験では，2 要因を分けて頭部動作の変化を計測した．また，小さな声で発声すると頭部が動かないことも予備実験にて確認されていたため，Mouth Openness Condition では，大きな声のみ発声させた．

頭部動作は被験者の頭頂に取り付けた Inertial Measurement Unit(IMU) で計測した．被験者には口の形をはっきり作るように教示することで，母音に対する口の開き具合を統制した．

3.2 実験手順

各条件ごとに被験者には 2 回試行させた．1 回目は実験室での発声に馴化するために行った．また，身体動作を正しく計測できているかの確認も行った．すべての発声後に，発音する際に意識した姿勢がどのようなものかアンケートにて調査した．

3.3 実験結果

実験被験者は 11 人（男：6 人，女：5 人，平均年齢 22.0，標準分散 0.54）であった．そのうち男性被験者 1 人が正しく声の高さを発声できていなかったため解析から除いた．

Voice Pitch Condition の計測結果を図 2 に示す．縦軸は発声定常状態での首の角度を示す．高音，中音，低音を発声する際の首の角度を分散分析にかけたところ，有意差が認められた ($F(2, 18) = 12.843, p < 0.01$)．さらに，多重比較したところ，高音を発声する際に首の角度が最も上がり ($p < 0.05$)，低音を発声する際に最も下がることが

明らかとなった ($p < 0.05$)．すなわち，高音を発声する際は頭部をそらし，低音を発声する際は頭部を下げる傾向が認められた．

Mouth Openness Condition の計測結果を図 3 に示す．縦軸は発声に伴う首の角度の変化量を示す．この変化量は，発話開始前と発声定常状態での首の角度の差の絶対値で定義した．口を開いて発声する「あ」「え」「お」群と口を閉じて発声する「い」「う」群に分け，発声に伴う首の角度の変化量の大きさを比較したところ，口の開きを伴う発声条件のほうが有意に首を大きく動かすことが認められた (ウィルコクソンの順位和検定, $p < 0.05$)．

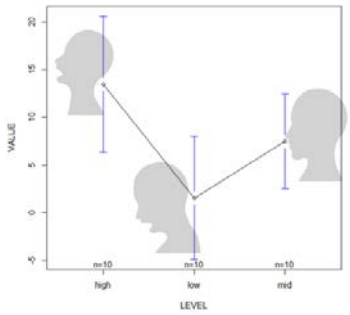

図 2: Head position according to pitch

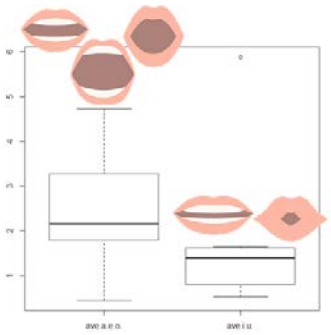

図 3: Head displacement according to mouth openness

以下にアンケートによる発声しやすい姿勢についての自由記述結果を示す．この記述からも，高音を発声する際は頭部をそらし，低音を発声する際は頭部を下げる傾向が認められた．

- 声の高低を意識して使い分けることが難しく感じ、高く出そうと思えば背筋が伸び顎が上がりました。低く出そうと思えば、背筋を少しだけ丸め顎を引き、なるべく口の中に篭るように発声しました。
- 高い音を出す際は上を向き、低い音を出す際には下を向く
- 口を大きく開ける あといは上から声を出し、うえおは下からあげるイメージで声を出す 体の中心に力を集めるイメージ
- 高い音は背筋が伸びる感じでした。低い音になるほど下を向いていたと思います。
- 高い音を出すときは顔を上向きに、逆に低い音を出すときは下向きにすると出しやすかった

4　身体的拘束に基づく発話動作生成システム

以上の知見をもとに音声特徴から頭部動作を生成するアルゴリズムを以下に説明する．人間らしい動作には滑らかな関節制御が重要である [25, 8]．そのため，音声特徴という間欠的な情報から連続的に滑らかな動作を生成する必要がある．また，二次遅れ系のダイナミクスに基づいて生成される動作が人間らしさ印象を与えることが報告されている [26]．そこで，本論文ではばねダンパ系を用いた運動モデルを利用することで，音声特徴という間欠的な情報から常時滑らかな動作を生成する（図 4, 式 1）．また，筋肉のモデル化をばねダンパ系を用いた運動モデルを用いた試みもあるため [27, 28]，この動作生成モデルの動作パラメータは筋肉の硬さに比例したパラメータとなっている．筋肉の硬さは発話の緊張度合・感情状態によって変化すると考えられ，発話時の感情や緊張度合といった人間が理解できるパラメータから動作パターンを調節することが期待される．

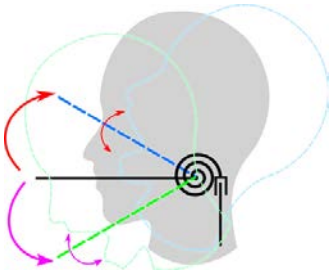

図 4: Classification of generating motion

$$J\ddot{\theta}_{base} + D\dot{\theta}_{base} + K\theta_{base} = T(t)Dir(t) \qquad (1)$$

4.1　ばねダンパ系による頭部動作生成

式 1 に対する外力を音声特徴をもとに定義することで，音声から頭部動作を自動生成する（式 2）．節 3 の実験結果から，口を大きく開けると首も大きく動くことから，式 4 のように，口の開く大きさによる外力を定義する．口の開きが大きくまたは均一である場合は，外力は口の開きの大きさに比例するようにする．口の開きが小さくなる場合は，首に与える外力をなくすことで運動モデルのばねの力により基準位置へ滑らかに戻る．口の開きが小さく

なる場合も口の開き度合をそのまま外力として与えてしまうと首の戻りが遅くなりリアルタイムで動作生成することが困難となる．また，予備実験から大きな声を出さないと首が顕著に動かなかったことから，声の大きさに比例した外力を式 3 のように外力を定義する．口の開き度合同様に，声のパワーが増えるまたは均一である場合は，外力は声の大きさに比例するようにする．声が小さくなる場合は，首に与える外力をなくすことで運動モデルのばねの力により基準位置へ滑らかに戻る．声が小さくなる場合も声のパワーをそのまま外力として与えてしまうと首の戻りが遅くなりリアルタイムで動作生成することが困難となる．V と L は声の大きさと口の開き度合という異なるスケールの外力を合わせるための定数である．

$$T(t) = VP(t) + LH(t) \tag{2}$$

$$P(t) = \begin{cases} Power(t) & (Power(t) >= Power(t-1)) \\ 0 & (otherwise) \end{cases} \tag{3}$$

$$H(t) = \begin{cases} LipHeight(t) & (LipHeight(t) >= LipHeight(t-1)) \\ 0 & (otherwise) \end{cases} \tag{4}$$

節 3 の実験結果から，首の動く方向は声の高さで決定されるため，式 1 の外力の運動モデルに対する方向を式 5 のように定義した．式 5 は，高音域を発声する場合は頭部をそらし，低音域を発声する場合は頷く方向に首を動かし，中音域では首を動かさないことを表す．

$$Dir(t) = \begin{cases} 1(Headup) & (HighTone) \\ -1(Headdown) & (LowTone) \\ 0(Nomovement) & (MiddleTone) \end{cases} \tag{5}$$

また，口の開閉度合は Ishi et.al. のフォルマント抽出に基づく口唇動作推定の手法を用いる [29].

4.2 韻律情報の抽出

F0 の値の抽出には，32 ms のフレーム幅で 10 ms 毎に LPC(Lear Predictive Coding) 逆フィルタによる残差波形の自己相関関数の最大ピークに基づいた処理を行う．さらに，人間のイントネーションの知覚特性と一致するよう，F0 の値を対数スケールに変換した．

$$F0[\text{semitone}] = 12 \times \log_2(F0[\text{Hz}]) \tag{6}$$

次に，音節内で F0 の変化量を表す $\Delta F0$(人間の音調の知覚に基づくパラメータ [30]) を抽出した．$F0move$ は音節の後半の F0 の近似直線上の音節末の F0($F0tgt2b$) と前半部の F0 平均値 ($F0avg2a$) との差分を用いて計算する（式 7）．そして，音節の音調は式 8 に応じて，上昇調，下降調，平坦調に分類した．

$$\Delta F0 = F0tgt2b - F0avg2a \tag{7}$$

$$tone = \begin{cases} rising\ (Rs) & (\Delta F0 > 1 \text{ semitone}) \\ falling\ (Fa) & (\Delta F0 < -2 \text{ semitones}) \\ flat\ (Ft) & (\text{otherwise}) \end{cases} \tag{8}$$

4.3 首と腰の協調動作

頭部が動く際には上下方向だけではなく，前後方向にも動くことが判っている [31]．このことから，首の 1 自由度の回転だけではなく，腰も連動させることでより人間らしい動きが実現できると考えられる．また，口と首の動き出すタイミングは異なり，口のほうがやや早いく動くことが報告されていることから [32]，動かす関節により位相差があることが考えらる．そこで，式 9 の変換式を用いて図 5 のような協調動作を実装する．

$$\theta_{act}(t) = \alpha_{act}\theta_{base}(t + \beta_{act}) \tag{9}$$

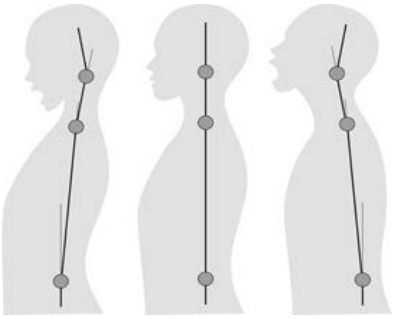

図 5: Multi-Joint Control

5 展望

節 4 で提案したモデルは，人間の身体的拘束に基づき，ばねダンパ系を用いた筋肉のダイナミクスを利用している．そのため，この動作生成モデルの動作パラメータは筋肉の硬さに比例したパラメータとなっている．筋肉の硬さは発話の緊張度合・感情状態によって変化すると考えられ，発話時の感情や緊張度合といった人間が理解できるパラメータから動作パターンを調節することが期待される．今後は，発話時の緊張・感情状態にあった動作を生成することができるかの検証や直感的に動作パラメータを決定できるかのユーザビリティの面から提案手法を評価する．

6 謝辞

本研究は，JST 戦略的創造研究推進事業 (ERATO） 石黒共生ヒューマンロボットインタラクションプロジェクトの一環として行われたものです．

参考文献

[1] Daisuke. Sakamoto, Takayuki Kanda, Tetsuo Ono, Hiroshi Ishiguro, and Norihiro Hagita. Android as a telecommunication medium with a human-like presence. In *Human-Robot Interaction*, pp. 193–200, 2007.

[2] 海光桑村, 竜二山崎, 修一西尾. テレノイドによる高齢者支援-特別養護老人ホームへの導入の経過報告-. 電子情報通信学会技術研究報告, Vol. 113, No. 272, pp. 23–28, 2013.

[3] Yutaka Kondo, Kentaro Takemura, Jun Takamatsu, and Tsukasa Ogasawara. A gesture-centric android system for multi-party human-robot interaction. *Journal of Human-Robot Interaction*, Vol. 2, No. 1, pp. 133–151, 2013.

[4] Miki Watanabe, Kohei Ogawa, and Hiroshi Ishiguro. Can Androids Be Salespeople in the Real World? In *ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 781–788, 2015.

[5] Masahiro Yoshikawa, Yoshio Matsumoto, Masahiko Sumitani, and Hiroshi Ishiguro. Development of an android robot for psychological support in medical and welfare fields. In *Robotics and Biomimetics*, pp. 2378–2383, 2011.

[6] Takuya Hashimoto and Hiroshi Kobayashi. Study on natural head motion in waiting state with receptionist robot SAYA that has human-like appearance. In *Robotic Intelligence in Informationally Structured Space*, pp. 93–98, 2009.

[7] Takanori Komatsu and Seiji Yamada. Adaptation gap hypothesis: How differences between users' expected and perceived agent functions affect their subjective impression. *Journal of Systemics, Cybernetics and Informatics*, Vol. 9, No. 1, pp. 67–74, 2011.

[8] Lukasz Piwek, Lawrie S McKay, and Frank E Pollick. Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition*, Vol. 130, No. 3, pp. 271–277, mar 2014.

[9] Binh Huy Le, Xiaohan Ma, and Zhigang Deng. Live Speech Driven Head-and-Eye Motion Generators. *Visualization and Computer Graphics*, Vol. 18, No. 11, pp. 1902–1914, 2012.

[10] Mehmet Emre Sargin, Yucel Yemez, Engin Erzin, and Ahmet Murat Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. In *Pattern Analysis and Machine Intelligence*, Vol. 30, pp. 1330–1345. Department of Electrical and Computer Engineering, University of California-Santa Barbara, Santa Barbara, CA 93106-9560, USA. msargin@ece.ucsb.edu, 2008.

[11] Carlos Busso, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation And Virtual Worlds*, Vol. 16, No. 3-4, pp. 283–290, 2005.

[12] Mary Ellen Foster and Jon Oberlander. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, Vol. 41, No. 3-4, pp. 305–323, 2007.

[13] Jelle Saldien, Bram Vanderborght, Kristof Goris, Michael Van Damme, and Dirk Lefeber. A motion system for social and animated robots. *International Journal of Advanced Robotic Systems*, Vol. 11, No. 1, pp. 1–13, 2014.

[14] Andrew G Brooks and Ronald C. Arkin. Behavioral overlays for non-verbal communication expression on a humanoid robot. *Autonomous Robots*, Vol. 22, No. 1, pp. 55–74, 2007.

[15] Miles L Patterson. 非言語コミュニケーションの統合モデルに向けて. 対人社会心理学研究, 第 7 巻, pp. 67–74, 2007.

[16] Tomio Watanabe, Masashi Okubo, Mutsuhiro Nakashige, and Ryusei Danbara. InterActor: Speech-Driven Embodied Interactive Actor. *International Journal of Human-Computer Interaction*, Vol. 17, No. 1, pp. 43–60, 2004.

[17] Chaoran Liu, Carlos Toshinori Ishi, H Ishiguro, and N Hagita. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Human-Robot Interaction*, pp. 285–292, 2012.

[18] Carlos Toshinori Ishi, ChaoRan Liu ChaoRan Liu, H Ishiguro, and N Hagita. Head motion during dialogue speech and nod timing control in humanoid

robots. In *Human-Robot Interaction*, pp. 293–300. Ieee, 2010.

[19] Kurima Sakai, Carlos Toshinori Ishi, Takashi Minato, and Hiroshi Ishiguro. Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *Robot and Human Interactive Communicationtion*, pp. 529–534, 2015.

[20] Per-Olof Eriksson, Hamayun Zafar, and Erik Nordh. Concomitant mandibular and head-neck movements during jaw opening-closing in man. *Journal of oral rehabilitation*, Vol. 25, No. 11, pp. 859–870, 1998.

[21] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, pp. 25–32, 2014.

[22] Randy J Larsen and Todd K Shackelford. Gaze avoidance: Personality and social judgments of people who avoid direct face-to-face contact. *Personality and Individual Differences*, Vol. 21, No. 6, pp. 907–917, 1996.

[23] Dwight Bolinger. *Intonation and Its Parts: Melody in Spoken English*. 1985.

[24] Hani C. Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, Vol. 30, No. 3, pp. 555–568, 2002.

[25] Michihiro Shimada and Hiroshi Ishiguro. Motion Behavior and its Influence on Human-likeness in an Android Robot. In *Annual meeting of the Cognitive Science Society*, pp. 2468–2473, 2008.

[26] 正幸中沢, 卓也西本, 茂樹嵯峨山. 力学モデル駆動による音声対話エージェントの動作生成. In *Human-Agent Interaction Symposium*, pp. 2C–1, 2009.

[27] Cho-chung Liang and Chi-feng Chiang. A study on biodynamic models of seated human subjects exposed to vertical vibration. *International Journal of Industrial Ergonomics*, Vol. 36, pp. 869–890, 2006.

[28] Astrid Linder. A new mathematical neck model for a low-velocity rear-end impact dummy: Evaluation of components influencing head kinematics. *Accident Analysis and Prevention*, Vol. 32, pp. 261–269, 2000.

[29] Carlos Toshinori Ishi, Chaoran Liu, Hiroshi Ishiguro, Norihiro Hagita, Intelligent Robotics, and Communication Labs. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. *IROS2012*, pp. 2377 – 2382, 2012.

[30] Carlos Toshinori Ishi. Perceptually-Related F0 Parameters for Automatic Classification of Phrase Final Tones. *IEICE transactions on information and systems*, Vol. 88, No. 3, pp. 481–488, March 2005.

[31] Hamayun Zafar, Erik Nordh, and Per-Olof Eriksson. Spatiotemporal consistency of human mandibular and head-neck movement trajectories during jaw opening-closing tasks. *Experimental Brain Research*, Vol. 146, No. 1, pp. 70–76, 2002.

[32] Hamayun Zafar, Erik Nordh, and Per-Olof Eriksson. Temporal coordination between mandibular and head-neck movements during jaw opening-closing tasks in man. *Archives of Oral Biology*, Vol. 45, No. 8, pp. 675–682, 2000.