I_039

マルチメディア信号におけるタイミング構造のモデル化

Modeling Timing Structure in Multimedia Signals

川嶋 宏彰† 堤 公孝† 松山 隆司†
Hiroaki Kawashima Kimitaka Tsutsumi Takashi Matsuyama

1. マルチメディア信号のタイミング構造

実世界で起こる様々なイベントを，複数のセンサで同時計測することで，複数のメディア信号が得られる．このとき，異なるメディア信号に現れる変化パターンの間には，共起性，時間的な同期などの時間的構造の特徴が見られ，計測したイベントの認識や生成にとって，しばしば重要な役割を持つ [8, 2].

マルチメディア信号における共起性のモデル化は，特徴量同士の相関を直接扱うものや，いったんHMMなどの状態を考え，異なるメディアの状態間での構造を扱うもの [3] がある．そして，これらの手法で注目している構造は，ほとんどの場合，同じ時刻や，隣接するフレームでの関係である（図1(a)）．しかし実際には，異なるメディアの変化パターン間に，これらのモデルでは直接表現しきれない構造が現れる．例えば，音声の破裂音/pa/と母音/a/を比べると，破裂音と唇動作の開始時刻はほぼ同期するのに対し，母音に対しては，唇の動きが若干先行することが多く，その開始時刻の時間差にはばらつきがある．また，ピアノなどの楽器の演奏では，実際の音に対して演奏者の手や体の準備的動作が入り，ライブ演奏においては豊かな情報を観客に与えている．

本論文では，このような異なるメディア信号の変化パターン間に起こる，共起性や系統的時間差といった時間的構造をタイミング構造と呼ぶ．このとき，マルチメディア信号におけるこのタイミング構造を明示的に表現する新たなモデル化の枠組みを示すことを目的とする．

個々のメディア信号は，信号内にあわられる有限個の要素的な変化(モード)によって表現できると仮定する．このような仮定の下で，メディア信号をモードの系列としてモデル化する手法としては，音声認識分野のsegment model [9] や，computer visionにおけるhybrid system [4, 5]，graphicsにおけるmotion texture [6] などが提案されており，物理的な時刻ではなく，状態やイベントの生起する時区間（以下では単に区間とする）とその遷移に基づいて複雑な変化を扱うことができる．ここでは，区間を単位としたこれらのモデルを，まとめて“interval model”と呼ぶことにする．

本論文では，あらかじめ個々のメディア信号が，それぞれinterval modelによって表現されていることを前提として，複数のinterval model間での時間的構造を，確率モデルによって表現する手法を提案する（図1(b)）．これにより，あるメディアの変化パターンと別のメディアの変化パターンがどの程度の時間差で始まり，どの程度の時間差で終了するかといった系統的な時間的関係を表現することが可能となり，さらに，あらかじめ学習された複数のメディア信号間のタイミング構造に基づいて，

†京都大学情報学研究科, Grad. Sch. of Informatics, Kyoto Univ.

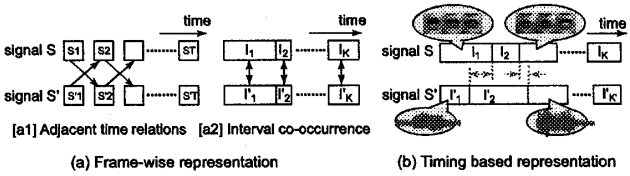


図 1: マルチメディア信号の時間構造のモデル化

認識や，メディア変換といった応用が可能になると期待される．本論文では，1つの応用例として，発話およびピアノ演奏における，音響・映像信号のタイミング構造モデルを学習し，これを用いることで音響信号から映像（唇や演奏者の動き）が生成できることを示す．

2. メディア信号間のタイミング構造モデル

2.1 メディア信号の時区間表現

マルチメディア信号におけるタイミング構造を定義するための前提として，個々のメディア信号は，それぞれinterval modelとして表現され，個々のinterval modelのパラメタは，あらかじめ学習されている（文献 [9, 6, 5] を参照のこと）とする．このとき，個々のメディア信号は，区間系列として以下のように記述することができる．

マルチメディア信号: マルチメディア信号とは，複数のセンサによって同時に計測して得られた N_s 個のメディア信号であるとする．このとき，メディア信号を S_c で表せば，マルチメディア信号は $\mathcal{S} = \{S_1, \cdots, S_{N_s}\}$ で定義することができる．各信号はそれぞれ $\Delta T_1, ..., \Delta T_{N_s}$ で標本化されているとする．

モードとモード集合: あるメディア信号 S_c の時間的な変化を表現するモード（要素的な変化）の集合を $\mathcal{M}^{(c)} = \{M_1^{(c)}, \cdots, M_{N_c}^{(c)}\}$ によって定義する．interval modelによって，モードのモデル化方法は異なり，文献 [4, 5] のように，個々のモードは線形システムによって表現されることが多い．

区間と区間系列: あるメディア信号 S_c が単一のモードに従って変化する時間範囲を区間 $I_k^{(c)}$ とする．区間 $I_k^{(c)}$ は開始時刻 $b_k^{(c)}$ と終了時刻 $e_k^{(c)}$，モードラベル $m_k^{(c)}$ をもつ．あるメディア信号 S_c を単一のモードに従う時間範囲に分割していくと区間系列 (区間の全順序集合) $\mathcal{I}^{(c)} = \{I_1^{(c)}, ..., I_{K_c}^{(c)}\}$ が得られる．なお，隣接区間 $I_{k-1}^{(c)}$ と $I_k^{(c)}$ は重なりを持たないとする．

マルチメディア信号の区間表現: 上記の記法を用いれば，マルチメディア信号 $\mathcal{S}$ の区間系列による表現は，集合 $\{\mathcal{I}^{(1)}, ..., \mathcal{I}^{(N_s)}\}$ となる．

2.2 タイミング構造の定義

本論文では2つのメディア信号 S, S' の間におけるタイミング構造に注目する（以下ではメディア信号を区別

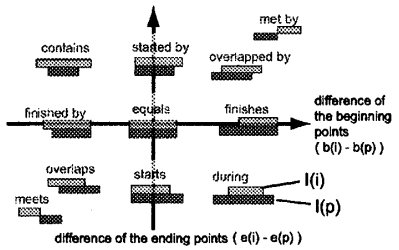


図 2: 始点や終点の時間差に基づく 2 つの区間の関係

するのに，' を用いる）．

メディア信号 S においてモード $M_i \in \mathcal{M}$ が現れるある区間を $I_{(i)}$ とする．（ここで，区間の生起する順序を表す k は今は重要ではないため省略した．）同様に，メディア信号 S' においてモード $M'_p \in \mathcal{M}'$ が表れるある区間を $I'_{(p)}$ とする．また，区間 $I_{(i)}$ における開始・終了時刻を $b_{(i)}, e_{(i)}$，同様に区間 $I'_{(p)}$ における開始・終了時刻を $b'_{(p)}, e'_{(p)}$ とする．これら 2 つのモードにおける時間関係は，一般にはこれら 4 つの時刻の関係 $R(b_{(i)}, e_{(i)}, b'_{(p)}, e'_{(p)})$ となる‡．本論文では，4 つの時刻の関係 R として，2 時刻間の二項関係 $R_{bb}(b_{(i)}, b'_{(p)})$, $R_{be}(b_{(i)}, e'_{(p)})$, $R_{eb}(e_{(i)}, b'_{(p)})$, $R_{ee}(e_{(i)}, e'_{(p)})$ の組み合わせによって定まるものを考え，これをモード間のタイミング構造と呼ぶ．

このとき，$R_{bb}, R_{be}, R_{eb}, R_{ee}$ として，具体的にどのような関係を考えるかが重要になる．それぞれに 3 つの時間的前後関係 $R_<, R_=, R_>$ を考えると 13 通りの関係になるが [1]，実際には，区間同士の時間的前後関係だけではなく，口の動きに対して，何 10msec 後に音声が生じたかといった，物理時間軸上での距離に基づくより詳細な関係が実用上重要となる．

そこで，まず時間的に離れたところにあるモードによる影響は小さいことを仮定する．そして，特に，区間同士が重なりを持つ，すなわち，R_{be}, R_{eb} をそれぞれ $R_\leq, R_\geq (b_{(i)} \leq e'_{(p)}$ かつ $e_{(i)} \geq b'_{(p)})$ とした場合における，R_{bb} と R_{ee} の時間差関係（始点差 $b_{(i)} - b'_{(p)}$ および終点差 $e_{(i)} - e'_{(p)}$）に注目する．すると，2 つの区間の時間関係は，図 2 に示すような 2 次元空間 $(b_{(i)} - b'_{(p)}) - (e_{(i)} - e'_{(p)})$ における点 $(D_b, D_e) \in \mathbb{R}^2$ となる．次節以降で提案するタイミング構造モデルは，このような重なりを持つモード間における開始時刻，終了時刻の時間差に注目してモデル化を行う点で，前後関係のみを扱う Temporal Interval Logic [1] に比べマルチメディアの同期現象を詳細に表現することができる．

2.3 タイミング構造のモデル化

重なりを持つモード対の時間差分布： 2 つのメディア信号 S, S' に，時間的な重なりを持って現れるモード対 $M_i \in \mathcal{M}, M'_q \in \mathcal{M}'$ において，その開始時刻の差 D_b および終了時刻の差 D_e の分布:

$$P(b_k - b'_{k'} = D_b, e_k - e'_{k'} = D_e | m_k = M_i, m'_{k'} = M'_p, [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \phi) \quad (1)$$

を，モード対の時間差分布と呼ぶ．この分布が原点付近に高い山を持つ場合は，その 2 つのモードは開始時刻および終了時刻が共に同期する傾向があるといえる．一方，系統的な時間差は現れない場合は，分布の分散が大きくなる．例えば，$b_k - b_{k'}$ は正の領域に鋭いピークを持ち，$e_k - e_{k'}$ の分散が非常に大きい場合は，常にモード M_i が M'_p に比べてほぼ同じ時間遅れて開始するが，終了時刻はどちらが先に終了するかも含めてばらつきを持つことになる．

モード対時間差分布は，与えられた区間系列から計算することができるが，サンプル数は有限個であるため，実用上は何らかの関数を当てはめることが望ましい．評価実験では 2 次元混合ガウス関数を用いた．

モード対の共起分布： 時間差分布は重なりをもつモード対に対する条件付き確率分布であり，どのモード対が重なる頻度が高いかという共起性を別途モデル化する必要がある．そこで，オーバラップを持つ区間対 $I_k, I'_{k'}$ において，それぞれのモード対が現れる頻度を，確率分布:

$$P(m_k = M_i, m_{k'} = M'_p | [b_k, e_k] \cap [b'_{k'}, e'_{k'}] \neq \phi). \quad (2)$$

で表現する．この共起行列は，重なりを持つ全ての区間対について，各モード対が現れる頻度分布を計算することで得られる．

モード遷移確率： 式 (1) および (2) を用いることで，2.2 節で定義したメディア間のタイミング構造を表現することができる．しかし，マルチメディア信号の特徴を記述するためには，タイミング構造だけでなく，モード間の順序関係を表現することも重要となる．ここでは，多くの interval model と同様に，モード間の遷移確率:

$$P(m_k = M_j | m_{k-1} = M_i) (M_i, M_j \in \mathcal{M}). \quad (3)$$

を用いて，隣接する区間のモードに現れる時間的順序関係の頻度をモデル化する．

3. タイミング構造に基づくメディア変換

発話や演奏などの動的イベントを，マイクやカメラなどの複数のセンサで同時に計測することで得られたマルチメディア信号から，2.3 節で導入したタイミング構造モデルをあらかじめ学習しておく．すると，入力された音響信号から，唇の動きや演奏者の動きを映像として生成をするといったメディア信号の変換が可能となる．本節ではその具体的な方法について述べる．

あるメディアの時系列信号 S' から別メディアの時系列信号 S を生成するメディア変換は以下の流れで実現できる: (1) 入力されたメディア信号 S' を，区間モデルに基づいて区間系列 $\mathcal{I}' = \{I'_1, ..., I'_{K'}\}$ へ分節化する．(2) メディア信号 S' の区間系列 $\mathcal{I}'$ から別のメディア信号 S

‡2 つのメディア信号の標本化レートが異なる場合，$b_{(i)}$ や $b'_{(p)}$ の代わりに $b_{(i)}\Delta T$ や $b'_{(p)}\Delta T'$ を用いて連続時間で構造を考える必要がある．本節ではこれも含め単に $b_{(i)}$ と表記し一般的な議論を行う．

区間系列 $\mathcal{I} = \{I_1, ..., I_K\}$ を生成する．(3) 生成された区間系列 $\mathcal{I}$ からメディア信号 $S : \{x_1, ..., x_T\}$ を生成する．なお，K, K' はそれぞれ区間系列 $\mathcal{I}, \mathcal{I}'$ に現れる区間の個数であり，一般には $K \neq K'$ である．

このうち (1),(3) については文献 [6, 5] で既に単一メディアのモデルを用いた方法によって実現されている．本論文ではこの手順 (2) における，一方のメディアの区間系列から他方のメディアへの区間系列の変換方法について提案する．以下では簡単のため，2つのメディア信号の標本化レートが一致しているものとする．

3.1 メディア変換の問題の定式化

2つのメディア信号間のあらかじめ学習されたタイミング構造モデルを Φ とする．このとき，一方のメディアの区間系列 $\mathcal{I}'$ を参照しながら，もう一方のメディアの区間系列 $\mathcal{I}$ を生成する問題を考える．これは，メディア信号 S' の区間系列 $\mathcal{I}'$ が与えられたときに，この区間系列と共に生じるもう一方のメディアの区間系列 $\mathcal{I}$ のうち，最も高い確率を取るものを見つけることで実現でき，以下のような最適化問題として定式化できる．

$$\hat{\mathcal{I}} = \arg\max_{\mathcal{I}} P(\mathcal{I}|\mathcal{I}', \Phi) \tag{4}$$

S' の長さを T とすれば，具体的には時間範囲 $[1, T]$ において，分節化を行う区間の個数 K，および各区間の終了（もしくは開始）時刻 $e_k(b_k)$ とそのモード $m_k (k = 1, ..., K)$ を決めることになる．区間の系列を決定する問題は非常に自由度が大きいため，全ての区間系列の候補 $\{\mathcal{I}\}$ についてそれぞれ式 (4) の確率値を計算して，最適な区間系列を求める方法では，T が大きくなるに従い膨大な計算が必要となる．そこで，HMM などにおける Viterbi アルゴリズムと同様に，動的計画法を用いて式 (4) の最適化問題を解く．なお，以降の式変形では Φ を省略する．

3.2 動的計画法による区間系列生成

時刻 t である区間が終了することを，文献 [7] の表記にならい $f_t = 1$ で表す．すると，区間系列 $\mathcal{I}'$ が与えられたときに，時刻 t においてモード M_j をとり，かつそこで区間が終了する確率は $P(m_t = M_j, f_t = 1|\mathcal{I}')$ と表せ，以下の漸化式

$$P(m_t = M_j, f_t = 1|\mathcal{I}') = \sum_{\tau} \sum_{p(\neq q)} \left\{ \begin{array}{l} P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} = M_i, \\ f_{t-\tau} = 1, \mathcal{I}') P(m_{t-\tau} = M_i, f_{t-\tau} = 1|\mathcal{I}') \end{array} \right\}$$

で計算することが可能である．ここで，l_t は時刻 t までに区間が持続している長さを，m_t は時刻 t におけるモードを表す．すると，ただちに以下の式に基づく動的計画法を導くことができる．

$$E_t(j) = \max_{\tau} \max_{i(\neq j)} P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} = M_i, f_{t-\tau} = 1, \mathcal{I}') E_{t-\tau}(i), \tag{5}$$

$$\text{where } E_t(j) \triangleq \max_{m_1^{t-1}} P(m_1^{t-1}, m_t = M_j, f_t = 1|\mathcal{I}')$$

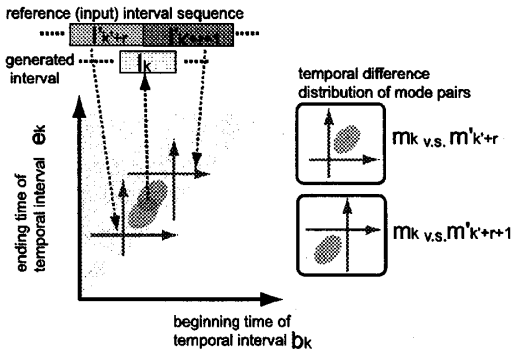


図 3: 区間遷移確率の計算の概要

ここで，$E_t(j)$ は，時刻 t においてモード M_j で区間が終了する確率の最大値であり，時刻1から $t-1$ の全ての可能なモード系列（パス）について最適化されている．

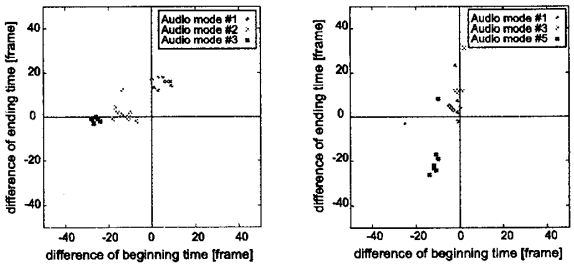
確率 $P(m_t = M_j, f_t = 1, l_t = \tau | f_{t-\tau} = 1, m_{t-\tau} = M_i, \mathcal{I}')$ は，参照区間系列 $\mathcal{I}'$ が与えられ，さらにモード M_i を持つ区間が $t-\tau$ で終了するときに，モード M_j を持つ区間が $[t-\tau+1, t]$ の範囲で現れる確率である．ここではこの確率を区間遷移確率と呼び，あらかじめ学習されたタイミング構造モデルおよびモードの遷移確率から計算する．紙面の都合上，具体的な導出方法を省略し，直感的な説明を図3に示す．まず，今注目している $[t-\tau+1, t]$ の区間に対して，オーバラップする参照区間（一般には複数）を見つける．すると，それら参照区間に対する生成区間の始点・終点の確率は，参照・生成区間の相対的な時間関係をモデル化した式 (1) を，式 (2) と組み合わせることで，絶対時間軸（生成区間の始・終点の2次元）にマッピングすることができる．この際，式 (3) によって直前の生成区間のモードを考慮することで，$[t-\tau+1, t]$ でモード M_j を取る確率が計算できる．

式 (5) により，各モードの区間が時刻 t で終了する確率の最大値を，時刻 $t = 1$ から $t = T$ まで再帰的に計算することができる．最後に，時刻 $t = T$ で終わる区間のうち最大の確率を取るモード $j^* = \arg\max_j E_T(j)$，および $E_T(j^*)$ を式 (5) で求めた際の最大値を与えた τ から，最後の区間のモードとその持続長が定まる．この操作を繰り返して trace back することにより，式 (4) の最適系列を得ることができる．区間の個数 K についても，この trace back を行った際に定まる．

4. 評価実験

音声から唇動画像を生成する実験を行い，提案手法の有効性を検証する．続いて，音響信号からピアノ演奏者のシルエット動画像の生成へも応用可能であることを示す．

特徴抽出: 人が母音/a//i//u//e//o/と連続して9回発話する様子（約18秒間）の映像を，解像度 720×480，フレームレート 60fps で撮影した．音声の標本化レートは 48kHz とした．その後，音声はフィルタバンク解析（フーリエ窓の幅は 1/30msec，窓間隔は 1/60msec）によって，映像は唇周辺の矩形領域を低解像度化し，得ら



(a) visual mode #1　　(b) visual mode #7

図 4: 音声，映像モード間の時間差の散布図．映像モード#1,7 はそれぞれ/o/ → /a/, /a/ → /i/の動きに対応

れた 32×32 の系列に主成分分析を行うことで，フレームレートを合わせた特徴ベクトル系列 1134 フレームを得た（音声は 25 次元，映像は 27 次元）．

各メディア信号の分節化とモード集合の推定:　音声と映像の特徴ベクトル系列をそれぞれ信号 S, S' と考え，それぞれのモード数，モードパラメタの推定，および分節化を行った．各モードとしては線形システムモデルを用いた．その方法として，線形システムの固有値制約に基づく階層的クラスタリング [5] を用いた．得られた各メディアの区間系列を図 5 の 1,2 段目に示す．それぞれの図では，縦方向にモードを並べて区間を表現した．横軸は時間方向である．音声と映像のモード数はそれぞれ 8 および 13 と自動推定された．音声のモードは画像に比べて多くなったが，これは同じ音が異なるモードになる場合があったためである．

音声と映像間のタイミング構造モデルの学習:　分節化によって得られた 2 つの区間系列を用いることで，式 (1)，(2)，(3) の確率分布を推定した．映像モードのうち 3 つについて，時間差の散布図を図 4 に示す．同じ母音でも，/a/への動きに比べ，/i/への動きの開始と音声はよく同期することが分かる．時間差分布はこれらに対して混合ガウス分布を当てはめることで推定した．

音声信号を入力とした唇映像生成:　学習に用いた音声区間系列に対して，画像区間系列を生成した結果を図 5 の 3 段目に示す．生成された画像区間系列と，あらかじめ学習された画像のモード（線形システム）集合を用いて，画像の特徴系列を生成した．その後，主成分分析における固有ベクトルとの線形和を計算し，各フレームの特徴ベクトルを画像化した．このうち，フレーム 140 から 250 までを，5 フレーム間隔で図 5 の 5 段目に示す．学習に用いた画像系列を 6 段目に示すが，両者の唇の動きはほぼ同期していることが分かる．さらに，同じ時間範囲における音声信号（図 5 の 5 段目）と比較すると，音声の開始に先行して唇が動くなど，タイミング構造モデルが詳細な時間構造を保持していることが分かる．

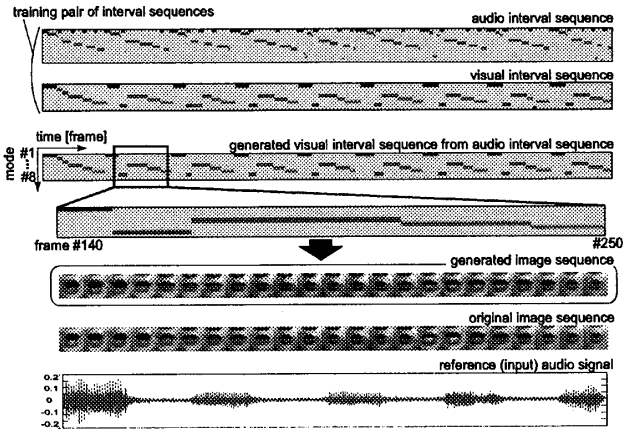


図 5: 音声信号を入力とする唇動画像の生成

5. おわりに

マルチメディア信号に含まれる各メディア信号は，その変化パターン同士に系統的な時間差や相互依存性，共起性性といった時間構造が存在すると考え，これらを区間に基づいてモデル化する手法を提案した．また，実際の音響信号から映像を生成する実験によって，その有効性を確認した．ただし，今回はモデルやアルゴリズムの基本的能力を検証する準備的な評価であり，学習時に用いた音響データそのものを入力して確認するに留まっている．未知のデータに対する生成へはそのまま適用可能であるが，より多くの実データに基づいた学習が必要になると考えられる．タイミング構造に基づく発話・イベント認識への展開も含め，今後の課題とする．

謝辞:　本研究の一部は，科学研究費補助金 18049046 の補助を受けて行った．

参考文献

[1] J. F. Allen. Maintaining knowledge about temporal interval. *Commun. of the ACM*, Vol. 26, No. 11, pp. 832–843, 1983.

[2] M. Brand. Voice puppetry. *Proc. SIGGRAPH*, pp. 21–28, 1999.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994–999, 1997.

[4] C. Bregler. Learning and recognizing human dynamics in video sequences. *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp. 568–574, 1997.

[5] H. Kawashima and T. Matsuyama. Multiphase learning for an interval-based hybrid dynamical system. *IEICE Trans. Fundamentals*, Vol. E88-A, No. 11, pp. 3022–3035, 2005.

[6] Y. Li, T. Wang, and H.-Y. Shum. Motion texture: A two-level statistical model for character motion synthesis. *Proc. SIGGRAPH*, pp. 465–472, 2002.

[7] K. P. Murphy. Hidden semi-Markov models (HSMMs). *Informal Notes*, 2002.

[8] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, Vol. 2002, No. 11, pp. 1–15, 2002.

[9] M. Ostendorf, V. Digalakis, and O. A. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Process*, Vol. 4, No. 5, pp. 360–378, 1996.