Evaluation de systèmes de génération de mouvements faciaux

O. Govokhina[1,2], G. Bailly[1], G. Breton[2] & P. Bagshaw[2]

(1) Institut de la Communication Parlée, UMR CNRS 5009, INPG/U. Stendhal, 46, av. Félix Viallet - F38031 Grenoble
(2) France Telecom R&D, 4 rue du Clos Courtel, BP 59, F35512 Cesson-Sévigné Cedex

{oxana.govokhina,gerard.bailly}@icp.inpg.fr, {gaspard.breton,paul.bagshaw}@francetelecom.fr

ABSTRACT

This paper presents the implementation and evaluation of different movement generation techniques for speech-related facial movements. State-of-the-art systems are implemented. A novel system that combines HMM-driven pre-selection of diphones with a standard concatenation system is also implemented. The trajectory formation systems are parameterised using the same training material. The ground-truth data consists of facial motion and acoustic signals of one female speaker uttering 238 sentences. Both objective and subjective evaluation of the systems is reported. The objective evaluation observes the linear correlation coefficient between original and predicted movements. It is complemented by an audiovisual preference test where ground-truth and predicted movements drive a 3D virtual clone of the original speaker.

I. INTRODUCTION

La perception et la production de la parole sont bimodales. L'information complémentaire et redondante fournie par l'articulation acoustique et visuelle est utilisée efficacement par les interlocuteurs pour améliorer la détection de la parole [1] et son intelligibilité [2]. Les personnes sont très sensibles aux divergences audiovisuelles spatiales [3] et temporelles [4]. Les systèmes de synthèse de la parole qui peuvent aussi produire des signaux audiovisuels à partir des données phonologiques ou acoustiques doivent reproduire les co-variations observées dans la parole naturelle ainsi que contrôler la variabilité des gestes articulatoires.
La modélisation de la coarticulation est un problème difficile et non résolu [5]. La variabilité de l'articulation observée est largement planifiée [6] et exploitée par les interlocuteurs [7]. Depuis les premiers travaux d'Öhman sur la modélisation des mouvements linguaux [8], plusieurs modèles de coarticulation ont été proposés. Nous avons implémenté quelques modèles et nous les avons paramétrés et comparés avec des données de capture de mouvement.
Cet article est organisé comme suit: l'état de l'art est brièvement présenté dans la section 2, le corpus audiovisuel utilisé et la modélisation articulatoire sont décrits dans la section 3, les systèmes de synthèse visuelle implémentés sont dans la section 4, la méthodologie d'évaluation et les résultats sont présentés dans les sections 5, 6 et 7.

II. ETAT DE L'ART

Trois composants essentiels constituent un système d'animation faciale : contrôle, forme et apparence. Le module de contrôle calcule un ensemble de paramètres caractéristiques de la forme à partir d'une spécification de la chaîne phonétique à prononcer. Le module de forme se charge de calculer une géométrie à partir des paramètres caractéristiques calculés puis le modèle d'apparence se charge de texturer cette forme géométrique. Les modules de contrôle peuvent être divisés en deux catégories en fonction du type des données d'entrée [9]: ils opèrent soit par mise en correspondance avec un signal corrélé i.e. acoustique, soit par calcul depuis une spécification symbolique i.e. la chaîne phonétique. Les systèmes guidés par l'acoustique essaient de générer les mouvements faciaux qui auraient produit les sons correspondants. Le décodage phonémique intermédiaire n'est pas obligatoire [10, 11].
Les systèmes opérant à partir de la chaîne phonétique peuvent être grossièrement divisés en différentes catégories: systèmes basés visemes [12], systèmes basés coarticulation [13, 14], systèmes de modélisation de trajectoires [15, 16] et systèmes basés concaténation [17-19].
La comparaison de ces systèmes est problématique [9] car les modèles sont construits à partir des différents corpus et leurs méthodologies d'évaluation sont différentes. L'approche modulaire est rarement possible car les modèles de contrôle, de forme et d'apparence sont souvent mélangés. La qualité du rendu aussi influence beaucoup la qualité des modèles de contrôle [20]. Ici, l'objectif est d'évaluer les systèmes de synthèse visuelle existants et de proposer un nouveau modèle basé données qui profite des meilleurs solutions existantes.

III. DONNEES AUDIOVISUELLES ET MODELISATION ARTICULATOIRE

Les systèmes de synthèse que nous allons évaluer sont construits à partir de données audiovisuelles. La base de données utilisée comprend 238 phrases du français prononcées par une locutrice. Les données acoustiques et vidéo sont capturées par un système Vicon© [21]. Le système capture, à 120 Hz, les positions 3D des 63 marqueurs infrarouges réfléchissants qui sont posés sur le visage de la lectrice (voir Figure 1). Le signal acoustique, à 11025 Hz, est segmenté semi-automatiquement en phonèmes.
Un modèle de forme est construit à partir des positions 3D des 63 points caractéristiques. La méthodologie du *clonage* développée à l'ICP [14, 22] consiste d'une Analyse itérative en Composants Principaux (ACP) appliquée sur des sous-ensembles pertinents des points caractéristiques. D'abord, les contributions de la rotation et de la protrusion de la mâchoire (*Jaw*1 and *Jaw*2) sont estimées et soustraites des données. Ensuite, Le mouvement d'arrondissement des lèvres (*Lips*1) est estimé à partir du résidu et soustrait des données. Les mouvements verticaux des lèvres du haut et du bas (*Lips*2 and *Lips*3), des coins des lèvres (*Lips*4) et de la gorge (*Lar*1) sont soustraits dans cet ordre des données résiduelles. Ainsi sept paramètres articulatoires sont obtenus. Leur contribution à l'explication la variance des mouvements est présentée dans le Tableau 1.

Tableau 1. Contribution des paramètres articulatoires à la variance globale.

Paramètre	*Jaw*1	*Lips*1	*Lips*2	*Lips*3	*Lips*4	*Jaw*2	*Lar*1
Variance	18.84	15.93	15.10	14.28	14.07	11.91	9.88

Cumulée	18.84	34.77	49.87	64.15	78.22	90.13	100

Un clone 3D est construit à partir des enregistrements vidéo photogrammétriques de la même lectrice [23]. Le clone virtuel vidéo-réaliste (voir Figure 1) est contrôlé par les sept paramètres articulatoires déterminés ci-dessous.

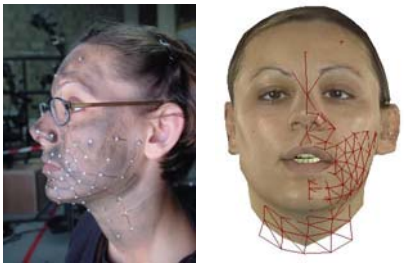

Figure 1. Gauche: Disposition des points caractéristiques utilisés pendant la capture des mouvements. Droite: Le clone virtuel 3D de la lectrice.

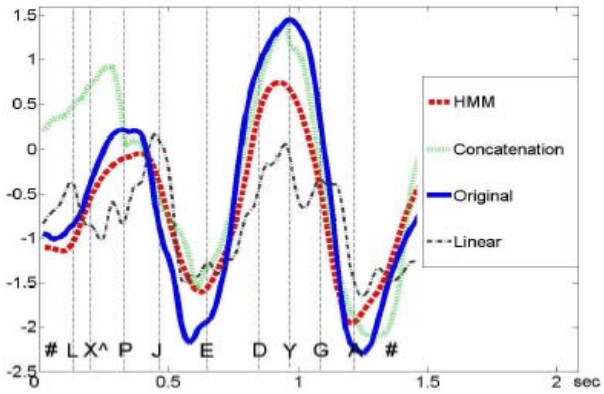

Figure 2. Trajectoires du paramètre JAW1 de rotation de mâchoire générée par différents systèmes pour la phrase "Le pied du gars".

Les systèmes de génération des mouvements faciaux liés à la parole sont paramétrés par les données audiovisuelles de 228 phrases. Les dix phrases restantes sont utilisées dans le test et contiennent des diphones qui peuvent être retrouvés dans la base des données en au moins deux exemplaires.

IV. LES SYSTEMES DE SYNTHESE VISUELLE

Plusieurs modèles de synthèse représentant les approches décrites auparavant sont implémentés:

IV.1. Le modèle linéaire guide par l'acoustique

Les paramètres articulatoires sont calculés directement à partir des paramètres acoustiques. Des signaux de 16.7 ms sont extraits à partir du signal d'origine en synchronie avec les données visuelles. Douze paramètres LSP (Line Spectrum Pair) et l'énergie sont calculés et lissés [24]. Enfin, un modèle de régression linéaire qui relie les paramètres acoustiques aux paramètres articulatoires est estimé. A la synthèse, les paramètres articulatoires sont générés à partir des paramètres acoustiques grâce au modèle obtenu.

IV.2. Le modèle de synthèse basé HMM

Le principe de synthèse vocale par HMMs fut introduit par Donovan [25] et étendu à la synthèse audiovisuelle par le groupe HTS [16]. La technique de synthèse par HMMs comprend les étapes d'apprentissage et de synthèse. Son application à la synthèse visuelle est décrite ci-dessous.

Apprentissage. Un HMM et un modèle des durées d'états sont appris pour les paramètres articulatoires de chaque diphone de la base d'apprentissage. Les vecteurs d'observation sont constitués des paramètres visuels statiques et dynamiques, c'est-à-dire, des valeurs des paramètres articulatoires et leurs dérivées. L'estimation des paramètres des HMMs est basée sur le calcul de maximum de vraisemblance (*Maximum-Likelihood criterion*) [26]. Cette estimation est effectuée par un algorithme spécifique de EM (*Expectation Maximisation*) connu comme l'algorithme récursive de Baum-Welch. Ainsi, un modèle gauche-droit à 3 états avec les distributions gaussiennes simples est appris pour chaque diphone.

Synthèse. La synthèse est effectuée comme suit. D'abord, la chaine phonétique à synthétiser est découpée en diphones (avec leurs durées respectives). Ensuite, une séquence des diphone-HMMsS correspondants est construite. Les durées des états sont déterminées [27]. Une fois la séquence d'états spécifiée, la trajectoire des paramètres articulatoires est estimée grâce à un algorithme spécifique de génération des paramètres [28]. Cet algorithme exploite la dépendance entre paramètres statiques et dynamiques. Ainsi, ce système est équipé théoriquement pour prendre en compte l'effet de coarticulation.

IV.3. Le modèle de synthèse par concaténation

La synthèse de la parole par concaténation consiste en la sélection et concaténation d'unités préenregistrées dans un dictionnaire. Tout d'abord des caractéristiques phonologiques sont utilisées pout sélectionner les unités candidates: la correspondance phonémique est demandée mais des contraintes phonotactiques (contexte phonémique, position dans la syllabe ou dans le mot) et/ou phonologiques de plus haut niveau peuvent être ajoutées [29]. Ensuite, un algorithme de programmation dynamique trouve un chemin optimal à travers le treillis des candidats qui minimise un coût cumulé de sélection et de concaténation.

Le coût de sélection incorpore souvent des pénalités dues aux contraintes de sélection non respectées. Si un modèle prosodique est disponible, la déviation entre paramètres prosodiques calculés et sélectionnés est aussi souvent prise en compte [30]. Le coût de concaténation est calculé en fonction de la distance entre les unités adjacentes à la frontière. Les caractéristiques statiques et dynamiques sont souvent considérées.

Ici, les candidats sont les diphones. Ils sont choisis en utilisant seulement les contextes gauche et droite. Aucun coût de sélection n'est considéré. Les coûts de concaténation sont égaux aux distances euclidiennes entre les paramètres articulatoires aux frontières des unités pondérées par la variance globale expliquée (voir Tableau 1).

Enfin, les trajectoires des unités sélectionnées sont élargies/compressées non linéairement pour correspondre aux durées des diphones puis un algorithme spécifique de lissage anticipatoire est appliqué [31].

IV.4. Le modèle de synthèse par concaténation basé HMM

Un nouveau modèle qui utilise la prédiction par HMMs pour présélectionner les candidats a été implémenté. Les diphones contextuels (tri-diphones) présélectionnés à la première étape

Mean correlation	Nat	Inv	Lin	HMM	Conc (N=3)
1	1,00	-1,00	0,17	**0,55**	0,50
2	1,00	-1,00	0,26	**0,63**	0,47
3	1,00	-1,00	0,26	**0,58**	0,30
4	1,00	-1,00	0,18	**0,70**	0,66
5	1,00	-1,00	0,41	0,56	**0,64**
6	1,00	-1,00	**0,62**	0,54	0,56
7	1,00	-1,00	0,12	**0,60**	0,41
8	1,00	-1,00	0,39	**0,55**	0,20
9	1,00	-1,00	0,33	0,49	**0,56**
10	1,00	-1,00	0,40	0,59	**0,67**
Global	1,00	-1,00	0,31	**0,58**	0,50

Tableau 2. Evaluation objective des modèles phrase par phrase.

Vote (%, Nb)	Nat	Inv	Lin	HMM	Conc (N=3)
1	14,30 (3)	4,80 (1)	0	14,30 (3)	**66,70 (14)**
2	33,30 (7)	0	0	28,60 (6)	**38,10 (8)**
3	19,00 (4)	0	0	**57,10 (12)**	23,80 (5)
4	28,60 (6)	0	0	**52,40 (11)**	19,00 (4)
5	**85,70 (18)**	0	0	9,50 (2)	4,80 (1)
6	0	0	0	38,10 (8)	**61,90 (13)**
7	**71,40 (15)**	0	0	28,60 (6)	0
8	**57,10 (12)**	0	0	38,10 (8)	4,80 (1)
9	**81,00 (17)**	0	0	14,30 (3)	4,80 (1)
10	38,10 (8)	0	0	**57,10 (12)**	4,80 (1)
Global	**42,9**	0,5	0	33,8	22,8

Tableau 3. Mean preference scores.

du système de concaténation sont ensuite classés dans l'ordre décroissant du coefficient de corrélation entre les trajectoires des diphones de la base des données et celles prédites par HMMs. Les N meilleurs candidats sont retenus dans le treillis pour la sélection finale du modèle de concaténation. Notons que N=∞ correspond au modèle de concaténation initial et qu'une méthode de sélection moins brutale aurait consisté à utiliser le coût de sélection pour pénaliser les segments les moins corrélés.

V. EVALUATION OBJECTIVE

Les modèles de synthèse visuelle proposés sont paramétrés à partir de la base d'apprentissage. Les dix phrases de test sont synthétisées. Le coefficient de corrélation linéaire (coefficient de Pearson) entre les trajectoires synthétiques et celles d'origine est utilisé pour l'évaluation objective.

Cette première évaluation est mise à profit pour paramétrer de manière optimale les systèmes.

Les corrélations moyennes dans le cas de la synthèse par HMMs augmentent si les paramètres dynamiques sont pris en compte pendant les phases d'apprentissage et de synthèse. La corrélation est significativement plus importante quand la dérivée première est utilisée. L'utilisation de la dérivée seconde n'augmente cette corrélation que de manière marginale.

La corrélation moyenne dans le cas de la synthèse par concaténation en fonction des différentes valeurs de N atteint une valeur optimale pour N=3.

VI. EVALUATION SUBJECTIVE

Le but du test subjectif utilisé est d'évaluer la préférence globale des modèles proposés par rapport aux mouvements faciaux d'origine. Il faut noter que cette référence – souvent absente dans l'ensemble des stimuli utilisés dans les tests publiés – est très importante [31, 32].

Les trajectoires articulatoires des dix phrases sont générées par trois modèles: (a) le système de synthèse basé HMM avec les vecteurs articulatoires comprenant la dérivée première (HMM); (b) le système de synthèse par concaténation avec la méthode de présélection proposée et N=3 (Conc); (c) le système de synthèse par modèle de régression linéaire (Lin). Cet ensemble est complété par les trajectoires originales (Nat) et leurs inverses (Inv) où les paramètres originaux sont multipliés par -1 de manière à fournir aux sujets une gamme assez large de qualité.

Un exemple de trajectoires générées est montré Figure 2. Les résultats de l'évaluation objective correspondant aux modèles retenus sont dans le Tableau 2. La corrélation moyenne est maximale pour la synthèse par HMMs. Dans le cas d'une phrase, la corrélation est plus importante pour le modèle linéaire que pour le modèle de concaténation.

Les paramètres articulatoires générés sont utilisés pour l'animation du clone virtuel de la lectrice (voir Figure 1). Le signal acoustique original est joué en synchronie avec les mouvements faciaux. Ici, le test de préférence moyenne (*Mean Preference Score: MPS*) est utilisé. Chaque participant doit alors choisir la séquence qu'il préfère parmi cinq pour chaque phrase. Les 21 sujets qui ont participé à l'expérience n'ont aucune pathologie audiovisuelle. Les sujets peuvent jouer les stimuli tant de fois qu'ils désirent et peuvent changer leurs choix. L'ordre initial des séquences pour chaque phrase est aléatoire. Le test est effectué dans un environnement de luminance contrôlé. Les conditions de la luminance de fond sont basées sur la ITU-R BT.500-9 (ITU-R, 1998).

VII. RESULTATS ET DISCUSSION

Les résultats du test subjectif sont dans le Tableau 3. Le modèle le plus préféré est l'original (42.9%) suivi par le modèle HMM (33.8%) et le modèle de concaténation (22.9%). Les scores de préférence pour les modèles linéaire et inverse sont très bas, 0% et 0.5% respectivement.

La méthode de synthèse par HMM est jugée comparable aux mouvements originaux; les phrases générées par HMM étant de plus toujours préférées par au moins deux personnes. La synthèse par concaténation guidée HMMs est moins performante mais les résultats dépendent des phrases. Il est intéressant de constater que les mouvements de synthèse (HMM ou concaténation) sont préférés aux originaux pour six des dix phrases. Cela peut provenir des imperfections des modèles de forme et d'apparence mais les mouvements générés par ces deux modèles de prédiction sont jugés globalement comme adéquats aux mouvements originaux.

Le modèle linéaire a le score le plus bas (voir aussi les résultats précédents obtenus par Gibert et al [31]) même si sa corrélation objective est parfois importante et même proche de celle obtenue par le modèle de concaténation pour certaines phrases.

VIII. CONCLUSIONS

Des différentes méthodes de synthèse visuelle sont évaluées objectivement et subjectivement. Une nouvelle méthode proposée concatène les segments articulatoires présélectionnés grâce à une méthode basée HMMs. L'utilisation de cette méthode augmente considérablement la

corrélation entre les trajectoires synthétiques et originales. Ce gain ne permet pas cependant d'atteindre ceux de la synthèse purement HMM. Dans l'ensemble, les résultats de l'évaluation objective sont confirmés par l'évaluation subjective. Le système HMM semble être le plus efficace et le mieux accepté.

L'étude des résultats montre cependant que les résultats des évaluations dépendent du contenu phonétique des phrases. Le modèle HMM, s'il est meilleur en moyenne partout, génère des trajectoires moins coarticulées que celles produites par le système par concaténation. C'est dans cet esprit que nous avions décidé de coupler la solide charpente construite par HMM avec la richesse des détails phonétiques capturés par la synthèse par concaténation. Nous allons continuer à suivre cette idée qui devrait à terme produire un système à la fois robuste et fin.

BIBLIOGRAPHIE

[1] K. W. Grant and P. F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *Journal of the Acoustical Society of America*, vol. 108, pp. 1197-1208, 2000.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215, 1954.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[4] N. F. Dixon and L. Spitz, "The detection of audiovisual desynchrony," *Perception*, vol. 9, pp. 719-721, 1980.

[5] W. J. Hardcastle and N. Hewlett, *Coarticulation: Theory, Data, and Techniques*. Cambridge, UK: Press Syndicate of the University of Cambridge, 1999.

[6] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, pp. 3-35, 1990.

[7] K. G. Munhall and Y. Tohkura, "Audiovisual gating and the time course of speech perception," *Journal of the Acoustical Society of America*, vol. 104, pp. 530-539, 1998.

[8] S. E. G. Öhman, "Numerical model of coarticulation," *Journal of the Acoustical Society of America*, vol. 41, pp. 310-320, 1967.

[9] J. Beskow, "Talking heads. Models and applications for multimodal speech synthesis." Stockholm: KTH, 2003, pp. 63.

[10] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," presented at 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, Kloster Seeon, Germany, 2000.

[11] S. Curinga, F. Lavagetto, and F. Vignoli, "Lips movements synthesis using time-delay neural networks," presented at EUSIPCO, Trieste - Italy, 1996.

[12] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," *International Journal of Computer Vision*, vol. 38, pp. 45-57, 2000.

[13] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Eds. Tokyo: Springer-Verlag, 1993, pp. 141-155.

[14] L. Revéret, G. Bailly, and P. Badin, "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," presented at International Conference on Speech and Language Processing, Beijing - China, 2000.

[15] T. Okadome, T. Kaburagi, and M. Honda, "Articulatory movement formation by kinematic triphone model," presented at IEEE International Conference on Systems Man and Cybernetics, Tokyo, Japan, 1999.

[16] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," presented at EUROSPEECH, Budapest, Hungary, 1999.

[17] S. Minnis and A. P. Breen, "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis," presented at International Conference on Speech and Language Processing, Beijing, China, 1998.

[18] O. Engwall, "Evaluation of a system for concatenative articulatory visual speech synthesis," presented at International Conference on Speech and Language Processing, Boulder - Colorado, 2002.

[19] F. J. Huang, H. P. Graf, and E. Cosatto, "Triphone-based unit selection for concatenative visual speech synthesis," presented at International Conference on Acoustics, Speech and Signal Processing, Orlando, FL, 2002.

[20] I. Pandzic, J. Ostermann, and D. Millen, "Users evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330-340, 1999.

[21] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun, "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech," *Journal of Acoustical Society of America*, vol. 118, pp. 1144-1153, 2005.

[22] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533-553, 2002.

[23] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, pp. 331-346, 2003.

[24] H. C. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23-43, 1998.

[25] R. Donovan, "Trainable speech synthesis," in *Univ. Eng. Dept.* Cambridge, UK: University of Cambridge, 1996, pp. 164.

[26] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[27] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," presented at International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[28] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into hmm-based speech synthesis," presented at ISCA Speech Synthesis Workshop, Pittsburgh, PE, 2004.

[29] P. Taylor and A. W. Black, "Speech synthesis by phonological structure matching," presented at EuroSpeech, Budapest, Hungary, 1999.

[30] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," presented at International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA, 1996.

[31] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," presented at IEEE Workshop on Speech Synthesis, Santa Monica, CA, 2002.

[32] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," Massachusetts Institute of Technology, Cambridge, MA, CBCL Paper #224/AI Memo #2003-003 February 2003.