Visual Prosody: Facial Movements Accompanying Speech

Hans Peter Graf, Eric Cosatto, Volker Strom, Fu Jie Huang
AT&T Labs Research, 200 Laurel Ave. South, Middletown, NJ 07748
{hpg,eric,vst,jhuangfu}@research.att.com

Abstract

As we articulate speech, we usually move the head and exhibit various facial expressions. This visual aspect of speech aids understanding and helps communicating additional information, such as the speaker's mood. In this paper we analyze quantitatively head and facial movements that accompany speech and investigate how they relate to the text's prosodic structure.

We recorded several hours of speech and measured the locations of the speakers' main facial features as well as their head poses. The text was evaluated with a prosody prediction tool, identifying phrase boundaries and pitch accents. Characteristic for most speakers are simple motion patterns that are repeatedly applied in synchrony with the main prosodic events. Direction and strength of head movements vary widely from one speaker to another, yet their timing is typically well synchronized with the spoken text.

Understanding quantitatively the correlations between head movements and spoken text is important for synthesizing photo-realistic talking heads. Talking heads appear much more engaging when they exhibit realistic motion patterns.

1. Introduction

Speech is usually accompanied by head movements, facial expressions, and gestures, applied by the speaker for underlining the meaning of the text. They can aid the understanding of the spoken text, but they also convey a lot of additional information about the speaker, such as the emotional state or the speaker's temper.

Nonverbal components in face-to-face communication have been studied extensively, mainly by psychologists. Such studies typically link head and facial movements or gestures qualitatively to parts of the text. Many of the more prominent movements are clearly related to the content of spoken text or to the situation at hand. For example, much of the body language in conversations is used to facilitate turn taking. Other movements are applied to emphasize a point of view. Some movements serve basic biological needs, such as blinking to wet the eyes. Moreover, people always tend to move slightly to relax some muscles while others contract. Being completely still is unnatural for humans and requires considerable concentration.

Beside movements that are obviously related to the meaning of the text, many facial expressions and head shifts are tied more to the text's syntactic and prosodic structure. For example, a stress on a word is often accompanied by a nod of the head. A rising voice at the end of a phrase may be underlined with a rise of the head, possibly combined with rising eyebrows. These are the type of facial and head movements we investigate in this paper. Since they are analogous to prosody in speech analysis, we call them visual prosody.

Little information exists about prosodic movements, and, to our knowledge, no quantitative results have been published that show how such head and facial movements correlate with elements of speech. Prosodic nods are mentioned sometimes in connection with animating faces, e.g. in [2][7][9], but few details are given. Eckman and Friesen studied extensively emotional expressions of faces [5] and also describe non-emotional facial movements that mark syntactic elements of sentences, in particular endings. But the emphasis is on head movements that are semantically driven, such as nods indicating agreement.

Our main interest in head and facial movements is to synthesize naturally looking talking heads. With sample-based animation techniques, articulation of speech can be emulated so realistically that synthesized heads are often hard to distinguish from recorded videos [3]. However, this is the case only for short sentences. In longer sequences a lack of naturally appearing head movements give the talking head a synthetic quality.

Many of the classical animation techniques have only limited applicability for the types of talking heads we describe here. Artists have been able since a long time to express emotions and personality in cartoon characters with just a few strokes of a pen [4]. However, as talking heads look more and more like real, recorded humans, viewers become more critical of small deviations from what is considered natural. For example, a cartoon character needs only very rough lip-

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

IEEE COMPUTER SOCIETY

sound synchronization to be perceived as pleasant. A photo-realistic head, on the other hand, has to show perfect synchronization. Otherwise the depicted person may seem to have a speech disability, which may be embarrassing to a viewer.

Similarly, the movements of a photo-realistic head have to be 'natural' or else viewers tend to dislike it. Sequences where, for example, the head moves randomly look very synthetic. The head seems to float over the background, which is judged by most viewers as 'eerie'. A talking head looks already much more natural if recorded head movements are added, even if they are not related with the text. Yet, truly satisfactory results are obtained only if head and facial movements are synchronized with the text.

Emulating human behavior perfectly requires an understanding of the content of the text. For unconstrained text this is beyond the capabilities of present-day natural language understanding programs. However, since many of the movements are closely coupled to prosodic elements of the text, we can derive naturally looking head movements using just the prosodic information.

2. Prosody

Prosody describes the way speech is intonated with such elements as pauses, pitch, timing effects, and loudness. The details of the intonation are influenced by the personality of the speaker, by the emotional state, as well as by the content of the text. Yet, underneath personal variations lie well-defined rules that govern the intonation of a language (see e.g.[8]). Predicting the prosody from text is one of the major tasks for text-to-speech synthesizers. Therefore, fairly reliable tools exist for this task.

The text we recorded was selected from four different sources:

- Short sentences, such as the one shown in table 3, plus greetings.
- Sentences designed to cover all diphones in English.
- Short children's stories.
- Paragraphs of the Wall Street Journal.

Overall this database contains 1,075 sentences. We recorded six different speakers. Five of them were talking for about 15 minutes each, pronouncing text from the first two sources. The sixth speaker was recorded for over two hours, articulating the whole data set. In this latter case the speaker was also instructed to speak some of the text while expressing a number of different emotions.

Our prosodic prediction tool identified prosodic phrase boundaries and pitch accents on the whole database, i.e. labeled the expected prosody. These events are labeled according to the ToBI (Tones and Break Indices) prosody classification scheme [1]. ToBI labels do not only denote accents and boundaries, but also associate them with a symbolic description of the pitch movement in their vicinity. The symbols shown in Table 1 indicate whether the fundamental frequency (F0) is rising or falling. The two tone levels, high (H) and low (L), describe the pitch relative to the local pitch range.

Table 1: ToBI symbols for marking pitch accents and phrase boundaries.

Symbol of pitch accent	Movement of the pitch of the fundamental frequency (F0)
H*	High - upper end of the pitch range.
!H*	Down-stepped high; somewhat lower than H.
L+H*	Low, moving high.
L*	Low - lower end of pitch range

Phrase boundary	Movements of F0
H-H%	Pitch high and rising higher towards end; typical for yes-no question.
L-H%	Pitch low and rising towards end; typical for comma.
L-L%	Pitch low, staying low; typical for end of a statement.

Table 2: Number of pitch accents and phrase boundaries in a data set of 1075 sentences.

Pitch accent	Number of events
H*	7277
L*	29
!H*	367
L+H*	608
Phrase boundary	
L-L%	664
H*, L-L%	712
L-H%	252
H*, L-H%	249
H-H%	9
H*, H-H%	13

Accents within spoken text are prime candidates for placing prominent head movements. Hence, their reliable identification is of main interest here. Stress within isolated words has been compiled in lexica for many different languages. Within continuous speech, however, the accents are not necessarily placed at the location of the lexical stress. Context or the desire to

IEEE COMPUTER SOCIETY

highlight specific parts of a sentence may shift the place of an accent. It is therefore necessary to consider the whole sentence in order to predict where accents will appear. Table 2 shows the different types of accents identified, and how many times they are encountered in the database.

Table 3: Phonetic transcript with prosodic annotation of the sentence: "I'm your virtual secretary". Three pitch accents appear in the text, and the sentence ends with a low pitch, which is typical for a statement.

Time (end of phone)	Phone	Prosodic event	Word
0.63 s	SIL		silence
0.77 s	ay		
0.85 s	m	H*	"I'm"
0.90 s	y		
1.00 s	er	none	your
1.05 s	v		
1.19 s	er	H*	
1.28 s	ch		
1.33 s	uw	none	
1.36 s	uh		
1.41 s	l	none	virtual
1.52 s	s		
1.61 s	eh		
1.67 s	k	H*	
1.71 s	r		
1.74 s	ih	none	
1.84 s	t		
1.94 s	eh		
2.06 s	r	none	
2.20 s	iy	L-L%	secretary

Any interruption of the speech flow is another event predestined for placing head or facial movements. Many disfluencies in speech are unpredictable events, such as a speaker's hesitations. 'ah' or 'uh' are often inserted spontaneously into the flow of speech. However, other short interruptions are predictably placed at phrase boundaries. Prosodic phrases, which are meaningful units, make it easier for the listener to follow. That is why prosodic phrase boundaries often coincide with major syntactic boundaries and punctuation marks. Table 2 shows the types of boundaries predicted by the prosody tool, and how often they appear in the text. With each phrase boundary a specific type of pitch movement is associated. This is of special interest here since it allows, for example, adding a rise of the head to a rising pitch. Such synchronizations can give a talking head the appearance of actually 'understanding' the text.

Table 3 illustrates the phonetic transcription and prosodic annotation of the text. In this case the phone durations were extracted from the spoken text with a phone labeling tool. Alternatively, the prosody analysis tools can predict phone durations from the text. Accents are shown here at the height of the last phone of a syllable, but it has to be understood that the syllable as a whole is considered accented and not an individual phone.

Of the different accent types, the H* accents strongly dominate (compare Table 2). Moreover, the prediction of the other types of accents is not very reliable. Even experienced human labelers agree in less than 60% on the accent type [10]. We, therefore, often do not differentiate between the various types of pitch accents and lump them all together simply as accents.

The prosody predictor has been trained with ToBI hand labels for 1,477 utterances of one speaker. The accent yes/no decision is correct in 89% of all syllables and the yes/no decision for phrase boundaries in 95%. The accent types are predicted correctly in 59% of all syllables and the boundary types in 74% of all cases. All the speakers recorded here are different from the speaker used to train the prosody predictor. We would expect the prosody prediction to be more accurate, if the tools could be trained on the same voice. But the achieved accuracies are actually quite close to the inter-labeler agreement for the same speaker [10] (a prosody LABELER is trained on the same voice as it is tested with).

3. Locating facial features

We are interested in natural head and facial movements. Hence, the speakers must be able to move their heads freely while they pronounce the text. Any invasive technique, such as placing sensors on the person's head was therefore ruled out. No markers or other artifacts for aiding the recognition system are used. We rely exclusively on the natural features in the face.

All the recordings were done with the speaker sitting in front of a teleprompter, looking straight into the camera. The frame size is 720 x 480 pixels and the head's height is typically about 2/3 of the frame height. The total of the recordings corresponds to 3 hours and 15 minutes of text. From these videos we extracted facial features and head poses for each of the over 700,000 frames (recordings were done at 60 frames per second).

In order to determine the precise head movements as well as the movements of facial parts, we have to measure the positions of several facial feature points with a high accuracy. Our face recognition system proceeds in multiple steps, each one refining the precision of the previous step [5]. Using motion, color

IEEE COMPUTER SOCIETY

and shape information, the head's position and the location of the main facial features are determined first with a low accuracy. Then, smaller areas are searched with a set of matched filters, in order to identify specific feature points with high precision. Figure 1 shows an example of this process. Representative samples of feature points, such as eye corners and edges of the eyes are cut from images. By averaging three of these images and band-pass filtering the result, the kernels become less sensitive to the appearance in one particular image. A set of such kernels is generated to cover the appearances of the feature points in all different situations. For example, nine different instances of each mouth corner are recorded, covering three different widths and three different heights of the mouth.

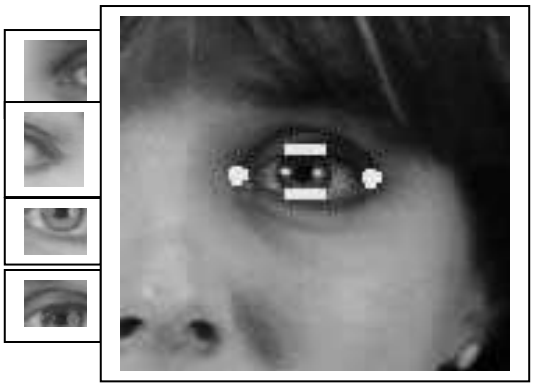

Figure 1: Locating parts of the eye. The kernels used for identifying eye corners plus the upper and lower edges are shown on the left side.

When a new image has to be analyzed, the first steps of the face recognition, namely shape and color analysis, provide information, how wide open the mouth is, and one can therefore select kernels of mouth corners corresponding to a mouth of similar proportions. Image and kernels are Fourier transformed for the convolution, which is computationally more efficient for larger kernels. In this way a whole set of filter kernels is scanned over the image, identifying the feature points.

The head pose is calculated from the location of the eye corners and the nostrils in the image. Figure 2 shows an example of identified feature points in the image and a synthetic face model in the same pose. Under the conditions of our setup the accuracy of the feature points must be better than one pixel; otherwise the resulting head pose may be off by more than one degree, and the measurements become too noisy for a reliable analysis.

There is a tradeoff between accuracy and selectivity of the filters. Larger filter kernels tend to be more accurate, yet they are more selective. For example, when the head rotates, the more selective filters are useable over a smaller range of orientations. Hence, more different filters have to be prepared. We typically tuned the filters to provide an average precision of between one and one and a half pixels. Then the positions are filtered over time to improve the accuracy to better than one pixel. Some events, for example eye blinks, can be so rapid that a filtering over time distorts the measurements too much. Such events are marked and the pose calculation is suspended for a few frames.

Figure 2: Feature points identified in the face. The head pose is calculated from the eye corners and the nostrils. The 3-d face model is shown in the same pose as the recorded face.

Beside the head pose we focus on the positions of the eyebrows, the shape of the eyes and the direction of gaze. These facial parts move extensively during speech and are a major part of any visual expression of a speaker's face. They are measured with similar filters as described above. They do not need to be measured with the same precision as the features used for measuring head poses. Whether eyebrows move up one pixel more or less does not change the face's appearance markedly.

The first part of our face analysis, where the head and facial parts are measured with a low accuracy, works well for any face. Sufficient redundancy is built into the system to handle even glasses and beards. The filters for measuring feature positions with high accuracy, on the other hand, are designed specifically for each person, using samples of that person's face.

4. Visual Prosody

For identifying prosodic movements, the rotation angles of the head around the x-, y-, and z-axis are determined, together with the translations. Figure 3 shows the orientation of the coordinate system used for these measurements. All the recorded head and facial movements were added spontaneously by the speakers while they were reading from the teleprompter. The speakers were not aware that the head movements

IEEE COMPUTER SOCIETY

would be analyzed. For most of the recordings the speakers were asked to show a 'neutral' emotional state.

For the analysis, each of the six signals representing rotations and translations of the head are split into two frequency bands:

- 0 – 2 Hz: Slow head movements
- 2Hz – 15Hz: Faster head movements associated with speech.

Movements in the low frequency range extend over several syllables and often over multiple words. Such movements tend to be caused by a change of posture by the speaker, rather than being related to the speech.

Figure 3: Orientation of the coordinate system. *ax, ay, az* mark the rotations around the x, y, z axes, respectively.

The faster movements, on the other hand, are closely related to the prosody of the text. Accents are often underlined with nods that extend typically over two to four phones. This pattern is clearly visible in Figure 4a. Here the nods are very clearly synchronized with the pitch accents (positive values for angle ***ax*** correspond to down movements of the head). Typical for visual prosody, and something observed for most speakers, is that the same motion - in this case a nod - is repeated several times. Not only are such motion patterns repeated within a sentence, but often over an extended period of time; sometimes over as much as whole recording session, lasting about half an hour.

A further characteristic feature of visual prosody is the initial head movement, leading into a speech segment after a pause. In Figure 4a this is shown as a slight down movement of the head (ax slightly positive), followed by an upward nod at the start of the sentence. We recorded 50 sentences of the same type of greetings and short expression in one recording session. The speaker whose record is shown in Figure 4 executed the same initial motion pattern in over 70% of these sentences.

In Figure 4 only the rotation around the x-axis, ***ax***, is shown. In this recording the rotation ***ax***, i.e. nods, was by far the strongest signal. Many speakers emphasize nods, but rotations around the y-axis are quite common as well, while significant rotations around the z-axis are rare. A combination of ***ax*** and ***ay***, which leads to diagonal head movements, is also observed often.

The mechanics for rotations around each of the three axes are different and, consequently, the details of the motion patterns vary somewhat. Yet, the main characteristics of all three of these rotations are similar and can be summarized with three basic patterns:

1. Nod, i.e. an abrupt swing of the head with a similarly abrupt motion back.
2. Nod with an overshoot at the return, i.e. the pattern looks like an 'S' lying on its side.
3. Abrupt swing of the head without the back motion. Sometimes the rotation moves slowly, barely visible, back to the original pose, sometimes it is followed by an abrupt motion back after some delay.

We can summarize these patterns with three symbols, where each one can be executed around the x-, y-, or z-axis:

- ^ nod (around one axis)
- ~ nod with overshoot
- / abrupt swing in one direction

Having such motion primitives allows describing head movements with the primitives' types, amplitudes and durations. This provides a simple framework for characterizing a wide variety of head movements with just a few numbers. Table 4 shows some statistical data of the appearance of these primitives in one part of the text database.

Table 4: Percent of pitch accents accompanied by a major head movement. Text corpus: 100 short sentences and greetings.

$P(^{}_x \mid *)$	42 %
$P(\sim_x \mid *)$	18 %
$P(/_x \mid *)$	20%

The amplitudes of the movements can vary substantially, as is illustrated by Figure 4b. For this recording the speaker was asked to articulate the same sentence as in Figure 4a, but with a cheerful expression. The initial head motion is now a wide down and up swing of the head, which runs over the first nod seen in Figure 4a. The first nod falls now on the second accent and the sentence ends with an up-down swing.

The patterns described here are not always visible as clearly as in the graphs of Figure 4. Some speakers show far fewer prosodic head movements than others. The type of text also influences prosodic head movements. When reading paragraphs from the Wall Street Journal, the head movements were typically less pronounced than for the greeting sentences. On the other hand, when speakers have to concentrate strongly, while reading a demanding text, they often exhibit very repetitive prosodic patterns.

IEEE COMPUTER SOCIETY

5. Conclusion

Head and facial movements during speech exhibit a wide variety of patterns that depend on personality, mood, content of the text being spoken, and other factors. Despite large variations from person to person, patterns of head and facial movements are strongly correlated with the prosodic structure of the text. Angles and amplitudes of the head movements vary widely, yet their timing shows surprising consistency. Similarly, rises of eyebrows are often placed at prosodic events, sometimes with head nods, at other times without. Visual prosody is not nearly as rigidly defined as acoustic prosody, but is clearly identifiable in the speech of most people.

Recent progress in face recognition enables an automatic registration of head and facial movements and opens the opportunity to analyze them quantitatively without any intrusive measuring devices. Such information is a key ingredient for further progress in synthesizing naturally looking talking heads. Lip-sound synchronization has reached a stage where most viewers judge it as natural. The next step of improvement lies in realistic behavioral patterns. We synthesized several sequences where the head movements consisted of concatenations of the motion primitives described above. With good motion-prosody synchronization the heads look much more engaging and even give the illusion that they 'understand' what they articulate.

All the investigations described here are done in the limited domain of a person reading from a teleprompter. A lot of work remains to be done to analyze behavioral patterns for spontaneous speech and during conversations.

6. References

[1] Beckman, M., Herschberg, J., "The ToBI Annotation Conventions", http://www.ling.ohio-state.edu/phonetics/ToBI/ToBI.6.html.

[2] Cassell, J, Sullivan, J. Prevost, S., Churchill, E., (eds.), "Embodied Conversational Agents", MIT Press, Cambridge, 2000.

[3] Cosatto, E. and Graf, H.P., "Photo-Realistic Talking-Heads from Image Samples", IEEE Trans. Multimedia, pp. 152-163, Sept. 2000.

[4] Culhane, S, "Animation; From Script to Screen", Martin's Press, New York, 1988.

[5] Ekman, P., Friesen, W.V., "Manual for the Facial Action coding system", Consulting Psychologists Press, Palo Alto, 1978.

[6] Graf, H. P., Cosatto, E. and Ezzat, T., "Face analysis for the synthesis of photo-realistic talking heads", Proc. *Fourth IEEE Int. Conf. Automatic Face and Gesture Recognition*, Grenoble, France, IEEE Computer Society, Los Alamitos, 2000, pp. 189-194.

[7] Hadar, U., Steiner, T.J., Grant, E.C., Rose, F.C., "The timing of shifts in head postures during converstion", Human Movement Science, 3, pp.237-245, 1984.

[8] Huang, X., Acero, A., Hon, H., "Spoken Language Processing", Prentice Hall, 2001, pp.739-791.

[9] Parke, F.I., Waters, K., "Computer Facial Animation", A.K. Peters, Wellesley, Massachusetts, 1997.

[10] Syrdal, A.K.,and McGory, J., "Inter-transcriber reliability and ToBI prosodic labeling", ICSLP 2000, Beijing, China; vol. 3, pp.235-238.

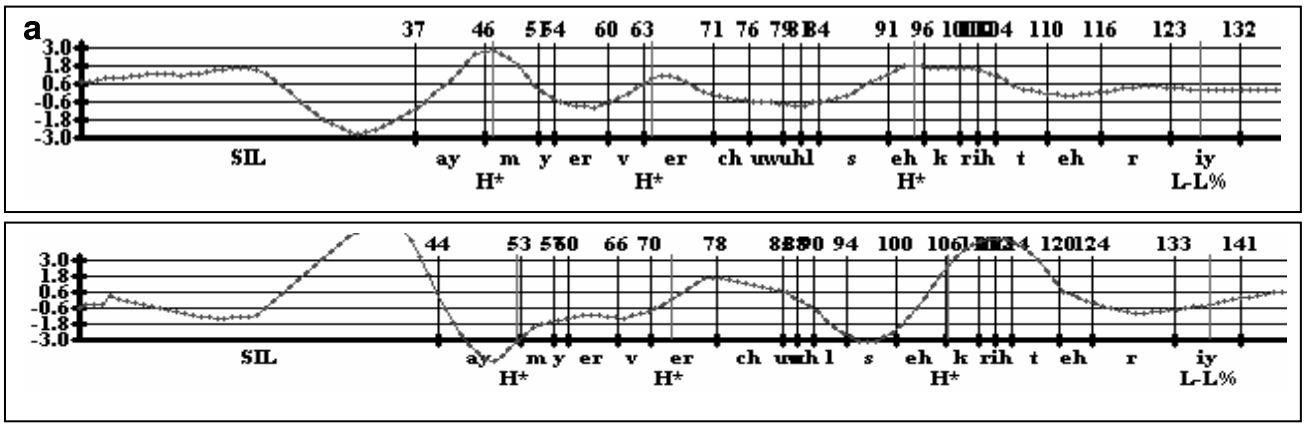

Figure 4a, b: Example of the head angle ax as a function of time. The vertical axes show ax in degrees. The top image (a) represents the high-pass filtered signal. Phone boundaries with frame numbers are marked at the top of the graph. Phones and prosodic events are shown below the graph. The bottom graph (b) shows the same sentence, spoken by the same speaker as in (a), but she was instructed to talk with a cheerful expression.

IEEE COMPUTER SOCIETY