Comparison of Phoneme and Viseme Based Acoustic Units For Realistic Lip Animation

May 9, 2007

Outline

- Problem definition
- Speech driven viseme recognition
- Compared acoustic units
- Overview of training & testing of HMMs
- Objective comparison of acoustic units
- Conclusion

Problem Definition

We aim to generate lip animation fullfilling the following requirements:

- Natural looking
- Synchronous to prerecorded speech
- Speaker – independent
- Using no text

Basic lip animation unit: Viseme

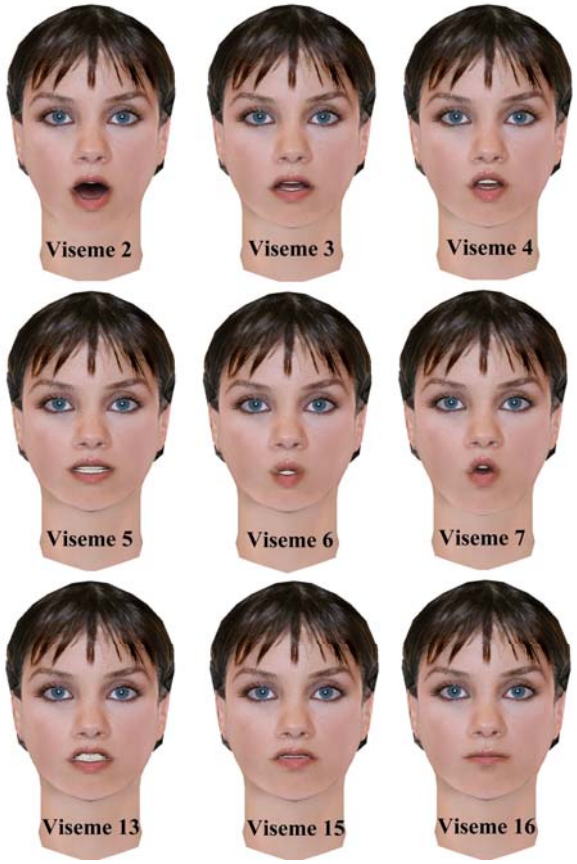

Figure 1. Example visemes for phoneme classes given in Table 1.

TABLE 1
PHONEME TO VISEME MAPPING

Viseme Classes	Timit Phoneset	Examples
1	pau	-
2	ay, ah	bite, but
3	ey, eh, ae	bait, bet, bat
4	er	bird
5	ix, iy, ih, ax, axr,y	debit, beet, bit, about, butter, yacht
6	uw, uh, w	boot, book, way
7	ao, aa, oy, ow	bought, bott, boy, boat
8	aw	bout
9	g, hh, k, ng	gay, hay, key, sing
10	r	ray
11	l, d, n, en, el, t	lay, day, noon, button, bottle, tea
12	s, z	sea, zone
13	ch, sh, jh, zh	choke, she, joke, azure
14	th, dh	thin, then
15	f, v	fin, van
16	m, em, b, p	mom, bottom, bee, pea

Compared Acoustic Units

- Phone based acoustic units
  - Phone HMM
  - Triphone HMM
- Viseme based acoustic units
  - Viseme HMM
  - Triviseme HMM

Strategy 1: Phone HMM

- Phone is the basic acoustic unit for human speech.
- There are 46 phones in the Timit phoneset.
- 46 phone-based HMMs.

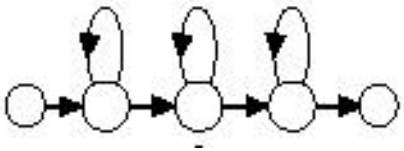

Figure 2. HMM structure with 5 states 3 of which are emitting states.

- Recognized phone sequences are mapped to viseme strings.
- This method does not consider effect of neighbour phones.

Strategy 2: Triphone HMM

- Triphone has a context – dependent structure.
- A phone with its left and right context makes a triphone (l – p + r).
- Recognized triphone sequences are mapped to visemes strings considering the center phone.

Strategy 2: Triphone HMM (ctd.)

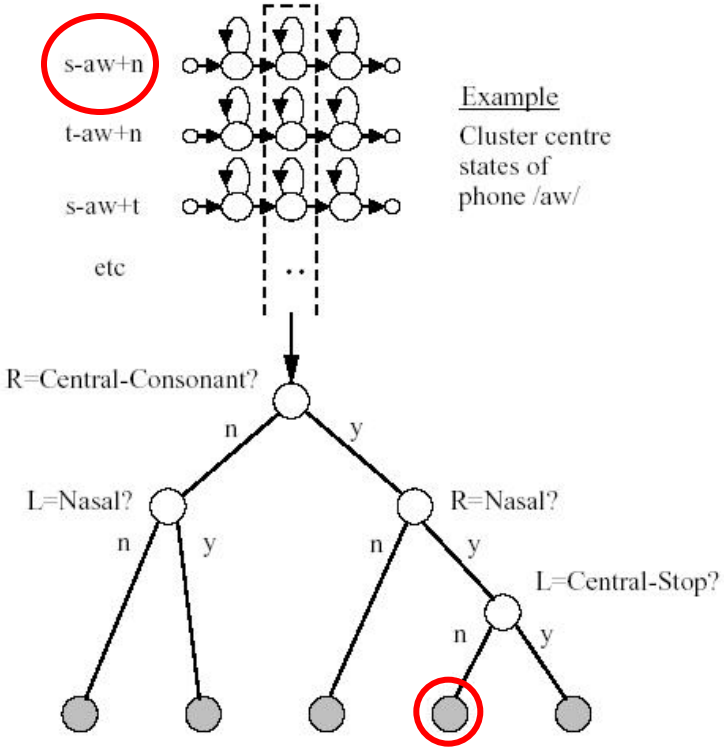

Figure 3. Decision tree-based state-tying.

- 11076 triphones.
- Decision - tree based state tying strategy.
- 202 questions.

Strategy 3: Viseme HMM

- Viseme: visual phoneme.
- We use 16 viseme classes.
- Timit has no viseme transcriptions. We use Table 1. to obtain viseme transcriptions.

Eg.	utterance:	ask		
	Phonetic transcription:	ae	s	k
	Viseme transcription:	3	12	9

Strategy 3: Viseme HMM (ctd.)

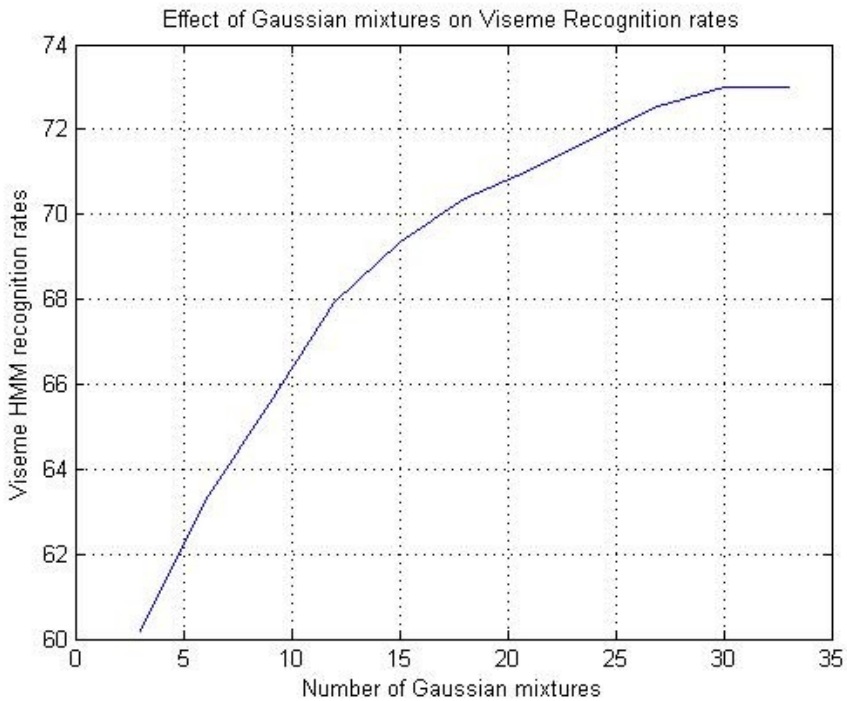

Figure 4. The plot of Gaussian mixtures versus viseme HMM recognition rates.

- Number of Gaussian mixture components is 30.
- Recognized viseme sequences have context – independent structure.
- No need for post – mapping recognition results since units are already visemes.

Strategy 4: Triviseme HMM

- Triviseme is a context – dependent structure: a viseme with left and right context makes a triviseme.
- There are 1941 trivisemes.
- Decision – tree based state tying strategy with 72 questions.
- Number of Gaussian mixture components is 6.
- Center viseme of a triviseme is used for lip animation, there is no need for post – mapping.

Training & Testing of HMMs

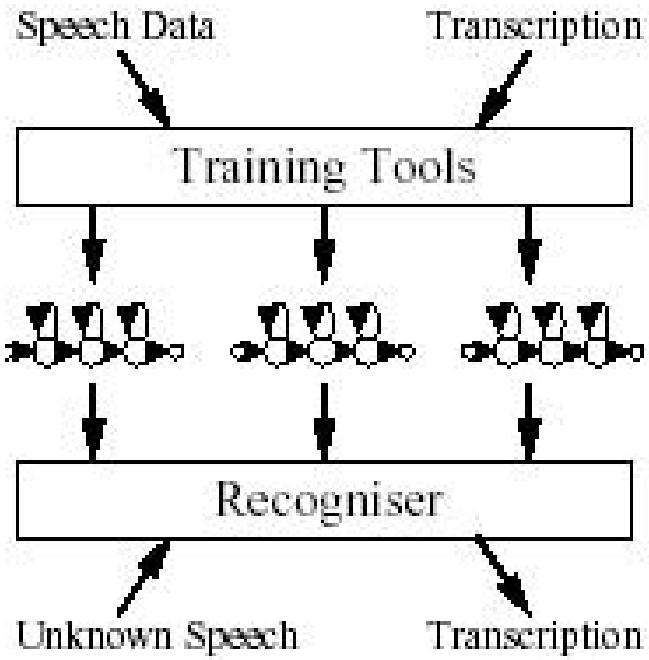

Figure 5. Training and recognition processes in HTK tool.

- Timit speech corpus
- HTK 3.1

Training of HMMs

- 462 training speakers, 3696 phonetically balanced training utterances.
- Training of HMMs is an incremental process.

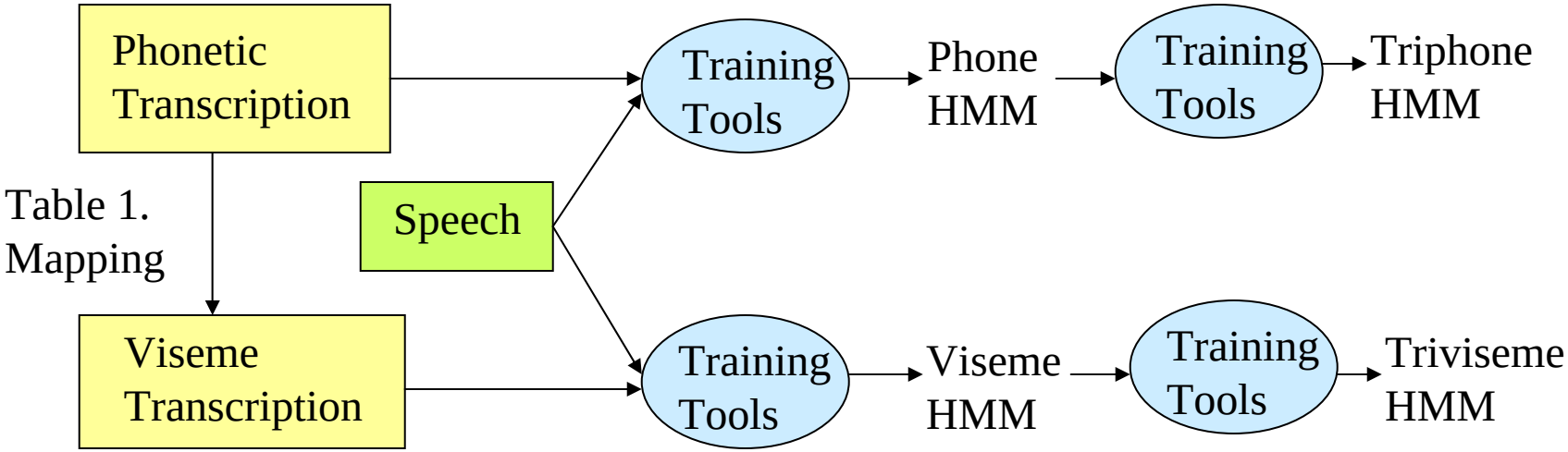

Figure 6. Training of HMMs

Testing of HMMs

- 168 test speakers, 1334 test utterances.
- Speaker –independent
- No need for transcription of the input speech

Objective Comparison of Acoustic Units

- TIMIT database

TABLE 2
VISEME RECOGNITION RATES FOR PHONE, TRI-PHONE, VISEME AND TRI-VISEME HMM MODELS

Method	Recognition Rate
Phone HMM	68.36 %
Tri-phone HMM	78.75 %
Viseme HMM	73.01 %
Tri-viseme HMM	79.49 %

Conclusion

- The performances of the phone, tri-phone, viseme and tri-viseme acoustic units are considered for HMM based viseme recognition. Based on the objective viseme recognition rates, we conclude that the tri-viseme based HMM structure outperforms the other structures.

References

[1] A.T. Erdem, “A New method for Generating 3D Face Models for Personalized User Interaction”, *13th European Signal Processing Conference*, Antalya, September 4-8, 2005.

[2] H. McGurk, and J. MacDonald, “Hearing Lips and Seeing Voices", *Nature*, vol 264, pp. 746-748, December 1976.

[3] S. Seneff, and V. Zu. “Transcription and Alignment of the Timit Database”, NIST, CD-ROM TIMIT, 1988.

[4] S. J. Young, D. Kershaw, J. Odell, and P.Woodland, “The HTK Book (for HTK Version 3.1)”, http://htk.eng.cam.ac.uk/, 2001.

[5] http://www.momentum-dmt.com/paper/tv_fdhc0.avi