AUDIOVISUAL SPEECH SYNTHESIS

Barry-John Theobald

School of Computing Sciences, University of East Anglia
bjt@cmp.uea.ac.uk

ABSTRACT

The ultimate goal of audiovisual speech synthesis is to create a machine that is able to articulate human-like audiovisual speech from text. There has been much interest in producing such a system over the last few decades and current state-of-the-art systems can generate very realistic synthesised speech. This paper presents a broad overview of audiovisual speech synthesis and considers possible future directions.

1. INTRODUCTION

Speech is a natural and efficient form of communication between humans. Speech signals are generally audiovisual and information rich, comprised of both linguistic information related directly to the message being conveyed and non-linguistic information related to such things as the identity and emotional state of the talker, the position in discourse, and so on. The ultimate goal of audiovisual text-to-speech synthesis (AVTTS) is to create a machine that is able to generate, from a textual string, expressive audiovisual speech that is indistinguishable from speech produced by a human.

AVTTS is extremely difficult because an orthographic representation of an utterance contains minimal information related to how an utterance should look and sound when spoken. Properties such as expression, speaking rate and prosody are undefined in the text, yet these are features a human reader will adjust naturally whilst reading text aloud. Likewise, the written form of a word is not a phonetic transcription of the spoken word. For example, the pronunciation of *receipt* is not reflected in the spelling. In addition, during natural speech production the effects of neighbouring speech gestures influence one another — a phenomenon known as *coarticulation* — and a synthesiser must infer these effects from the text.

Speech synthesis systems strive for *realistic* synthesis. A fundamental definition of realism would relate only to intelligibility: a listener should be able to understand the words and phrases and the visual gestures must be congruent with the auditory speech. Modern systems must consider higher level interpretations of realism. For example, the relative timing of the speech segments and the expressiveness, pleasantness and friendliness must all be perceived as being natural. A simplistic model of speech might consider linguistic information to be expressionless and non-linguistic features added to, or imposed upon, the underlying speech. However, it is likely the relationship is more complex and to date several models have been proposed [39]. A difficulty in improving the realism of synthesised speech is *naturalness* is subjective. A viewer/listener might easily be able to detect inaccuracies in synthesised audiovisual speech, but may not necessarily be able to identify *what* exactly is wrong in some measurable sense that allows refinement/improvement of the synthesis parameters.

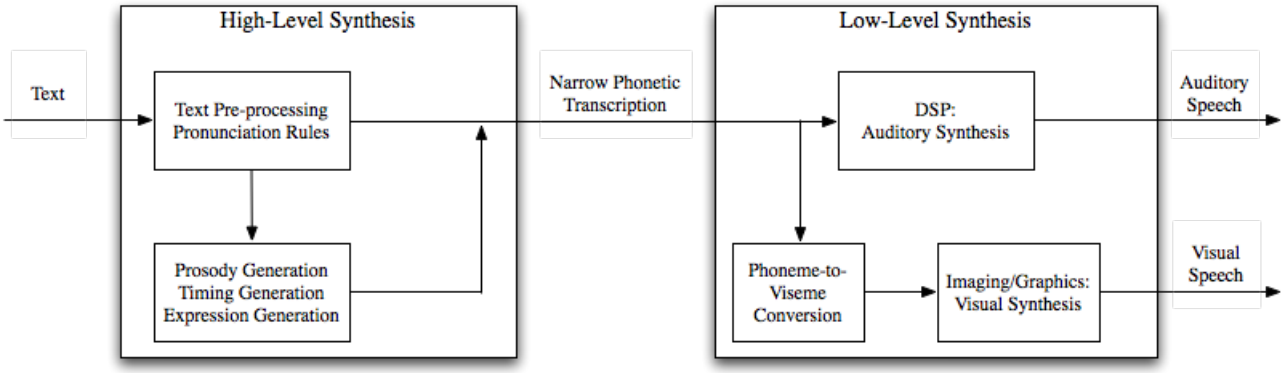
2. AUDIOVISUAL SYNTHESIS

Broadly speaking text-to-speech synthesis is a two-stage process: high-level synthesis utilises natural language processing (NLP) to convert a text string to a suitable parameterisation for a low-level synthesiser, and low-level synthesis utilises signal and image processing to generate the auditory waveform and accompanying visual gestures. A block diagram of a generic AVTTS system is shown in Figure 1. The interface between the high- and low-level synthesis modules is referred as a *narrow phonetic transcription* [18] to signify it is not only a sequence of phonetic symbols, but also includes prosodic information, timing and possibly expressive information.

The term audiovisual speech synthesis is somewhat of a misnomer: audio *and* visual speech synthesis would perhaps be a better description. Generally, at the lower-levels of synthesis audio and visual speech are synthesised independently as the form of the output signal for each modality is different. There are, however, exceptions that do consider the auditory and visual information jointly [15, 23].

The following sections first briefly outline the main steps involved in high-level synthesis, before describing the main (low-level) techniques for generating synthesised audio and visual speech signals.

Figure 1: A simplistic overview of an AVTTS system.



2.1. High-level Synthesis

The goal of high-level synthesis is to pre-process a text string to derive features that characterise the utterance to be synthesised. First the string is parsed and acronyms and abbreviations are disambiguated and expanded into the words that are to be "spoken". For example, *Dr.* could represent either *doctor* or *drive*. To overcome this the parser inspects neighbouring words and uses a set of rules (grammar) to determine the appropriate expansion: if *Dr.* is followed by a capitalised word (likely a name) it should be taken to be *doctor* [28]. Each word must then be fully spelled-out in terms of its pronunciation, which is usually achieved by looking up the appropriate entry in a pronunciation dictionary. A relatively simple dictionary might contain only a list of words and their constituent phonemes, whereas more sophisticated dictionaries might also contain diacritics relating to the stress pattern of the syllables within the words. Entries with multiple pronunciations can be disambiguated using, for example, decision trees that take into account the context in which the word appears. Morphophonological rules are applied to expand word roots and adjust (isolated) pronunciations given the overall utterance structure.

The final stage of high-level synthesis is to add prosodic features that relate to timing, intonation, and stress over the entire utterance. Natural synthesised speech requires appropriate pauses, in terms of position and duration, which can be determined stochastically or using syntactical rules [36]. For tonal languages it is extremely important the appropriate pitch contour is generated for the syllables within the utterance. In such languages the same word can have a number of meanings, and the intended meaning is inferred from the pitch pattern across the syllables that form the word(s). For non-tonal languages, stress must be added to the appropriate syllables to ensure the synthesised utterance conveys the correct meaning. For example, stressing the first syllable in the word *present* implies a gift (noun) or that an object is not missing (adjective), whilst stressing the second syllable implies an offering (verb). If the incorrect stressing/pitch contour is used more effort is required on the part of the listener and in extreme cases an utterance may not be intelligible at all. The appropriate intonation is required to ensure the interpretation of the utterance is as intended. For example, raising the pitch toward the end of a sentence is used to signal a question. Modern synthesisers that consider expressive speech will also add information to denote the desired expression and the intensity of that expression. This will in turn impact on features such as the overall loudness and speaking rate, which tend to increase with excitement or anger for example.

The narrow phonetic transcription generated by the high-level synthesiser fully characterises the utterance and is input to the low-level synthesiser for generating the acoustic waveform and accompanying facial gestures. For visual synthesis an additional mapping of auditory units (phonemes) to the visual counter-part (visemes) may be performed. This is usually just a simple look-up, although there are several phoneme-to-viseme mappings [32, 34].

2.2. Low-level Synthesis

Automatic computer-generation of synthesised auditory speech has been an active research topic since the mid 1960s, whereas work on automatic generation of visual speech began in earnest in the late 1980s–early 1990s. However, visual synthesis has benefited greatly from prior work on auditory synthesis, where techniques have been translated to the visual modality. Most visual synthesisers utilise the high-level component of an auditory synthesiser, so the visual synthesiser can be thought of as add-on

module to the TTS system. The following sections outline some of the main techniques for audio and visual synthesis.

2.2.1. *Articulatory Synthesis*

The aim of articulatory synthesis is to synthesise speech by simulating the biomechanics of speech production. Visual synthesis uses physically-based approaches [1, 3, 8, 20, 35] to simulate the interaction of the facial anatomy and auditory synthesis models the relationship between the change in position of the articulators and the corresponding change in transfer function of the vocal-tract system [4, 10, 38]. Articulatory synthesis has the promise of generating the most natural synthesised speech as it directly models the physical aspects of human speech production. However, it is the most difficult and least studied of the techniques as data capture and analysis require specialised equipment, e.g. electropalatography (EPG), X-Ray, magentic resonance imaging (MRI), or electromagnetic articulography (EMA) and the computational requirements are relatively high.

2.2.2. *Rule-based Synthesis*

Where articulatory synthesis attempts to model the *generation* of the audiovisual speech, rule-based synthesis is concerned only with modelling the end result — referred to as *terminal analogue* synthesis. Rule-based auditory synthesisers, often referred to as formant synthesisers [2, 24], parameterise short segments of real speech in terms of the average fundamental frequency, spectral components, and noise levels over the duration of a segment. Rule-based visual synthesisers [5, 7, 31] use parameters to control directly a model of the visual articulators (lips, teeth and tongue). A simplistic image-based approach represents speech gestures as static images and generates synthesised sequences by morphing between image pairs [22], or better still parameterises and re-synthesises visual speech by solving for combinations of morphs between a number of images [21]. Alternatively, graphics-based approaches control the position of virtual articulators during speech. One approach [37] implements a numerical model of coarticulation [33], while a more common approach [31] implements a gestural model [30]. The dominance and rate of movement of each articulator are defined for each visual speech gesture and overlapping exponential functions used to interpolate articulator positions between segments. This is an effective method and has undergone several refinements. For example, automatically estimating the dominance parameters for each gesture from real data [29], and improving the synthesis of bilabials and synthesising across languages [16]. A similar approach was also extended to image-based synthesis [13],

Rule-based synthesis has the advantages that storage of only minimal information is required, a single value for each parameter in each speech segment, and a potentially large number of sounds/gestures can be generated. The main limitations are realistic audiovisual speech is difficult to generate because the variability observed in natural speech must be re-introduced, without which the synthesised speech is perceived as being synthetic. Also, synthesis rules are (generally) derived manually using trial-and-error, which is time-consuming to develop.

2.2.3. *Concatentaive Synthesis*

Data-driven, concatenative speech synthesis does not synthesise speech directly. Rather novel phrases are synthesised by extracting, concatenating and normalising *units* of speech from a pre-recorded corpus [26]. There is no attempt to model the underlying properties of the audiovisual speech signal or its generation.

For auditory speech synthesis a cost function is designed to maintain smoothness in some spectral property of the acoustic signal across concatenation boundaries [17, 27, 41]. Generally, the best matching unit is considered the candidate that requires the *least* modification to form the join. For visual speech synthesis a cost function is designed to ensure a fluent and natural transition between adjacent visual speech gestures [9, 14, 25, 40]

Typical units used in concatenative audiovisual synthesis include phonemes, diphones, triphones, and demisyllables (and the visual counterparts visemes, disemes, trisemes, etc). Longer units minimise the number of concatenations and preserve coarticulation effects, but these require increasingly large corpus sizes to ensure a good coverage of transitions between the units. Consequently, long synthesis units are applied only in limited-domain applications. Shorter synthesis units require less storage and it is easier to ensure complete coverage of all transitions, but it is difficult to accurately segment the training utterances and to ensure prosodic features in the synthesised utterance are natural. To trade-off the disadvantages of longer versus shorter synthesis units, variable length units can be used and it is also possible to consider the joint audio *and* visual cost, as opposed to independent audio/visual costs [15, 23].

Concatenative synthesis is generally the most favoured technique and is used in many commercial speech synthesis systems. It is the most simple of the techniques as there is no attempt to model

the complexities of the audiovisual speech signal. In addition, concatenative synthesis generally produces the most natural output of the three synthesis strategies. The disadvantage of concatenative synthesis is a lack of flexibility. The synthesiser cannot generally extrapolate to instances of a speech unit not seen in the training data.

3. FUTURE DIRECTIONS

The following sections outline open issues and possible future directions for improving the naturalness of synthesised audiovisual speech.

3.1. Expressive Synthesised Audiovisual Speech

Computers are becoming more powerful and more affordable and storage devices are increasing in size, allowing increasingly large corpora of audiovisual speech to be collected, stored and searched. However, while current state-of-the-art synthesised audiovisual speech is intelligible, it still falls short of the naturalness of real speech. Human speech is not used only to articulate words. Rather it is a rich and expressive form of communication. Most audiovisual speech synthesis efforts to date have focussed on the concept of "neutral" speech and attempt to re-synthesis *speech* sounds that are without expression or emotional context. All synthesised speech, even that produced by state-of-the-art commercial systems, is identifiable as being synthesised. It is perhaps because this idea of neutral speech is itself unnatural, so despite best efforts, such systems will always fail to convince a listener/viewer that the speech is real.

To be more "human-like", AVTTS systems must be capable of altering the tone-of-voice and generating the respective facial gestures as required. Recent efforts capture separate recordings for the various expressions the synthesiser is to recreate, then either attempt to separate the speech and expression subspaces [12], or simply select samples from the appropriate recording as required [11]. However, these systems are limited to expressions that exist in the original recordings, which are usually a subset of the six primary expressions of emotion. Expressive speech is about more than just these basic expressions of emotion and a videorealistic synthesiser must be able to convey the full spectrum of expression beyond the basic emotions. Such expressions might include excitement, anticipation, boredom, disdain, fatigue, contempt, relief, and so on. In terms of scalability, the capture of a separate recording for *every* possible expression is clearly not feasible, so future efforts must focus on the integration/separation of expression from speech.

Recent work on automatic recognition of facial expression beyond the basic expressions of emotion [19] and automatic modelling of facial behaviours [6] perhaps offer some direction for future investigation. Expression-only subspaces could automatically be extracted from expressive speech information, and the evolution of expressions (behaviours) in this subspace learned from examples. Facial expressions, and other non-linguistic facial gestures, are then represented as template behaviours, which could be appended to a expressionless speech model. An issue with this is the capture of suitable data for training expressive speech synthesisers. Generally, information is lost during synthesis. So, to ensure what might be considered a natural level of expression in the synthesiser output, the training data may require overly expressive speech. The difficulty then is ensuring over-exaggerated expressive speech does not in itself degrade the naturalness of the synthesised expressive speech.

3.2. Speaker Independence

Concatenative audiovisual synthesisers are generally not only constrained to speech units in the training corpus, but also to the identity of the talker(s) in the original corpus. As speech synthesisers become more and more natural, we might expect more widespread use in multimodal interfaces. In this instance it will become increasingly important to allow the identity of the talker to be easily changed. For concatenative, data-driven, synthesisers this becomes difficult without re-recording the entire training corpus for each speaker of interest. The use of hybrid image- and graphics-based models (in the visual modality), such as those used to synthesise visual speech in [40], allow expressive visual speech information to be transferred between models, as illustrated in Figure 2. Ongoing experiments conducted between UEA/CMU/University of Virginia and University of Notre Dame, demonstrate expressive (visual) speech can be cloned in real-time. The quality of the cloned (non-synthesised) visual speech is sufficient that a number of viewers have been engaged in a real-time conversation, yet none have realised they are speaking to a cloned face. This has the advantage that a large corpus need only be collected for a single talker. Each subsequent identity need only capture a few tens of images from which to build their model. Visual speech is generated first on the original face, and later cloned to new faces before being displayed.

3.3. Reactive Interfaces

As the use of multimodal interfaces becomes more widespread we might expect interfaces to be *reac-*

Figure 2: Cloning of expressive visual speech information between (A) a source video and (B–C) two target faces..

(A) (B) (C)

tive. Rather than simply instructing a synthesiser to generate a phrase with a particular expression, the interface should detect the emotions, expressions, or the intentions of the user and modify the output accordingly. The interface should be aware of, and respond to, the surrounding environment. For example, if a system is able to detect a mistake on the part of a user it should be able to inform the user without causing annoyance or embarrassment. Without this the system would likely be counterproductive as the user would turn off this "helpful" functionality. Likewise, if the system detects annoyance or frustration on the part of the user it should have mechanisms for placating the user, or at least preventing the situation from becoming further inflamed — particularly if the cause of the frustration is the system itself. These issues go beyond basic audiovisual synthesis and require an understanding of how we interact with each other and with machines. For example, if a system is responsive and displays appropriate expressions, would we detect these as emotions and project a personality onto the system? Would the way we use computers change? Would this change improve productivity? The current state-of-the-art falls a long way short of this level of realism, and to achieve anything close in the future would require close collaboration with researchers in the fields of speech analysis, synthesis and perception, as well as wider fields, including behavioural psychology.

3.4. Standardisation

Progress towards fully-videorealistic audiovisual synthesis requires standardisation of experimental data, testing methodologies and experimental conditions. This is especially true for visual synthesis where there are currently no standards. In terms of data, a large, freely available audiovisual corpus is required so different systems can use the same data and their output can be fairly compared. Currently research groups tend to use their own data making direct comparisons difficult. Ensuring standard testing methodologies and test conditions will allow results obtained at different sites/times to be easily compared. Most (visual) synthesisers undergo little, or no, formal evaluation, so comparing results is meaningless as the training and test conditions are different.

4. SUMMARY

This paper has provided a brief overview of audiovisual speech synthesis and outlined some possible directions for future research. The ultimate goal of audiovisual speech synthesis is videorealism: that is, synthesised audiovisual speech that is indistinguishable from speech produced by a human. To achieve this an immediate concern must be the synthesis of *expressive* speech. Furthermore the expressiveness of the synthesiser must be over and above the expression of basic emotions.

For more widespread use of synthesisers, person-independent synthesis is desirable. It must be straightforward to change the identity of a talker, and this should not require the full capture and storage of a training database for each talker.

5. ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of EPSRC (EP/D0490751).

6. REFERENCES

[1] Albrecht, I., Schröder, M., Haber, J., Seidel, H. 2005. Mixed Feelings: Expression of non-basic emotions in a muscle-based talking head. *Journal of Virtual Reality*, 8(4):201–212.

[2] Allen, J., Hunnicut, S., Klatt, D. 1987. *From Text to Speech: The MITalk System*. Cambridge University Press.

[3] Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., Savariaux, C. 2002. Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3):533–553.

[4] Bangayan, P., Alwan, A., Narayanan, S. 1996. From MRI and acoustic data to articulatory synthesis: A case study of the laterals. In *Proceedings of International Conference on Spoken Language Processing*, pages 793–796.

[5] Beskow, J. 2003. *Talking heads: Models and applications for multimodal speech synthesis*. PhD Thesis, Centre for Speech Technology, KTH, Stockholm, Sweden.

[6] Bettinger, F., Cootes, T. 2004. A model of facial behaviour. *International Conference on Face and Gesture Recognition*, pages 123–128.

[7] Bevacqua, E., Pelachaud, C. 2004. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 15(3–4):297–304.

[8] Birkholz, P., Jackèl, D., Kröger, B. 2006. Construction and control of a three-dimensional vocal tract model. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 873-876.

[9] Bregler, C., Covell, M., Slaney, M. 1997. Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360.

[10] Brownman, C., Goldstein, L. 1986. *Towards an articulatory phonology*. In Ewan, C. and Anderson, J. (eds) *Phonology Yearbook 3*, Cambridge University Press, pages 219–253.

[11] Bulut, M., Narayanan, S., Syrdal, A. 2002. Expressive speech synthesis using a concatenative synthesizer. In *Proceedings of the International Conference on Spoken Language Processing*.

[12] Chuang, E., Bregler, C. 2005. Mood swings: Expressive speech animation. *ACM Transaction on Graphics*, 24(2): pages 331–347.

[13] Cosatto, E., Graf, H. 1998. Sample-based synthesis of photorealistic talking heads. In *Proceedings of Computer Animation*, pages 103–110.

[14] Cosatto, E., Graf, H. 2000. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia*, 2(3):152–163.

[15] Cosatto, E., Potamianos, G., Graf, H. 2000. Audio-visual unit selection for the synthesis of photo-realistic talking-heads. In *International Conference on Multimedia and Expo*, pages 619–622.

[16] Cosi, P., Magno Caldognetto, E., Perin, G., Zmarich, C. 2002. Labial coarticulation modeling for realistic facial animation. In *Proceedings of International Conference on Multimodal Interfaces*.

[17] Donovan, R. 2001. A new distance measure for costing spectral discontinuities in concatenative speech synthesisers. In *ISCA Tutorial and Research Workshop on Speech Synthesis*, pages 59–62.

[18] Dutoit, T. 1997. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers.

[19] el Kaliouby, R., Robinson, P. 2005. *Real-time inference of complex mental states from facial expressions and head gestures* in Real-time vision for HCI. Springer.

[20] Engwall, O. 2004. Speaker adaptation of a three-dimensional tongue model. In *Proceedings of International Conference on Spoken Language Processing*, pages 465–468.

[21] Ezzat, T., Geiger, G., Poggio T. 2002. Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH*, pages 388–398.

[22] Ezzat, T., Poggio T. 1998. Miketalk: A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference* pages 96–103.

[23] Fagel, S. 2006. Joint audio-visual unit selection — The JAVUS speech synthesizer. In *Proceedings of the International Conference on Speech and Computer*.

[24] Holmes, J., Mattingly, L., Shearme, J. 1964. Speech synthesis by rule. *Language and Speech*, 7:127–143.

[25] Huang, F., Cosatto, E., Graf., H. 2002. Triphone based unit selection for concatenative visual speech synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2037–2040.

[26] Hunt, A., Black, A. 1996. Unit selection in a concatenative speech synthesis system. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 373–376.

[27] Klabbers, E., Veldhuis, R. 2001. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9(1):39–51.

[28] Kleijn, W., Paliwal, K. 1995. *Speech coding and synthesis*. Elsevier Science.

[29] Le Goff, B., Benoît, C. 1996. A Text-to-audiovisual-speech synthesizer for French. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2163–2166.

[30] Löfqvist, A. 1989. Speech as audible gestures. In Hardcastle and Marchal (eds.) *Speech production and speech modelling*. Kluwer Academic Publishers, 289–322.

[31] Massaro, D. 1998. *Perceiving Talking Faces*. The MIT Press.

[32] Montgomery A., Jackson, P. 1983. Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America*, 73(6):2134–2144.

[33] Öhman, S. 1967. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41(2):310–320.

[34] Owens, E., Blazek, B. 1985. Visemes observed by the hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28:381–393.

[35] Pelachaud, C., Badler, N., Steedman, M. 1991. Linguistic issues in facial animation. In *Computer Animation '91*, pages 15–30.

[36] Read, I., Cox, S. 2005. Stochastic and syntactic techniques for predicting phrase breaks. In *Proceedings of the European Conference on Speech Communicaiton and Technology*.

[37] Reveret, L., Bailly, G., Badin, P. 2000. MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *Proceedings of the International Conference on Spoken Language Processing*, pages 755–758.

[38] Rubin, P., Baer, T., Mermelstein, P. 1981. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70: 321–328.

[39] Tatham, M., Morton, K. 2004. *Expression in speech: Analysis and synthesis*. Oxford University Press.

[40] Theobald, B., Bangham, J.A., Matthews, I., Cawley, G. 2004. Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, 44:127–140.

[41] Wouters, J., Macon, M. 1998. Perceptual evaluation of distance measures for concatenative synthesis. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2747–2750.