INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The Influence of Synthetic Voice on the Evaluation of a Virtual Character

João Paulo Cabral[1], *Benjamin R. Cowan*[2], *Katja Zibrek*[1], *Rachel McDonnell*[1]

[1]School of Computer Science & Statistics, Trinity College Dublin, Ireland
[2]School of Information & Communication Studies, University College Dublin, Ireland
cabralj@tcd.ie, benjamin.cowan@ucd.ie, zibrekka@tcd.ie, ramcdonn@tcd.ie

Abstract

Graphical realism and the naturalness of the voice used are important aspects to consider when designing a virtual agent or character. In this work, we evaluate how synthetic speech impacts people's perceptions of a rendered virtual character. Using a controlled experiment, we focus on the role that speech, in particular voice expressiveness in the form of personality, has on the assessment of voice level and character level perceptions. We found that people rated a real human voice as more expressive, understandable and likeable than the expressive synthetic voice we developed. Contrary to our expectations, we found that the voices did not have a significant impact on the character level judgments; people in the voice conditions did not significantly vary on their ratings of appeal, credibility, human-likeness and voice matching the character. The implications this has for character design and how this compares with previous work are discussed.

Index Terms: expressive speech, synthetic voice evaluation, avatars

1. Introduction

Virtual characters are popular in a wide range of domains including computer games, movies and customer services. When developing a virtual character, the graphical realism and voice used are significant design considerations, with developers regularly aiming for human-like graphics and speech. Speech synthesis research has recently focused on developing expressive synthesis so as to improve the naturalness of artificial voices [1, 2] ensuring that voices used by virtual characters and agents seem more human-like. As expressiveness of synthesis improves, synthesised speech may become a viable option for voicing of animated characters in audio-visual applications, such as computer games and films. That said, little is understood about the role that expressive synthetic speech plays in impacting people's judgments of virtual characters with which it is used. Our work aims to fill this gap by exploring how using expressive synthesis affects people's views of virtual characters.

Much of the current work around virtual character perceptions focuses on the impact of a character's human-likeness. The more anthropomorphic the representation of virtual characters is, the more competent and trustworthy they are perceived to be [3]. In terms of voice, recent qualitative research on intelligent personal assistants (IPAs) suggests that humanness of synthesis leads users to imbue IPAs with intelligence and personality [4, 5].

Research on the interplay between visuals and speech in character evaluation [6] shows that a mismatch between humanness of the voice and a character's graphical representation can negatively affect people's perceptions towards a character. Whenever people interacted with an avatar that had an audio-visual mismatch (e.g., either a virtual humanoid avatar with a human voice or a video of a human with a humanoid voice), they trusted the character less and had lower ratings of self-disclosure than when the character's visual representation matched the voice used. This audio-visual congruence is important when conveying expressivity in characters [10] because both channels tend to give variable levels of information to assess an emotion's valence and activation [10].

Although there seems to be an interaction between visuals and speech in character level judgments, speech tends to be a more significant driver of comprehension based judgments [6]. In a study observing people's perceptions of characters that recited book reviews, participants rated their understanding more highly when it was given in a human rather than humanoid voice, regardless of the character's graphical representation [6]. Voice, rather than both visuals and voice, seems to be the main dimension that people consider when making judgments about understanding of the content being delivered by the character.

In the speech synthesis field there is a growing need to test speech synthesis in context and consider a wider number of variables when evaluating synthetic voices. Current gold standard approaches focus on perceptual evaluations in which people rate the speech quality in terms of several aspects that include naturalness, intelligibility, intonation, prosody, voice pleasantness, rhythm and emotion. For example, the Blizzard Challenge [7] (`http://www.festvox.org/blizzard/`) is an annual Text-to-Speech Synthesis (TTS) shared task that performs this type of synthetic speech evaluation. More recently, the 2016 [8] and 2017 Blizzard Challenges have developed to consider speech synthesis from an expressive speech corpus related to a specific context (children's audiobooks).

Our work looks to evaluate the effect of expressive speech synthesis on both voice level and character level evaluations of a virtual character. We use the paradigm of personality as the dimension of expressivity, creating a synthetic voice based around the personality dimensions researched in previous work [11]. They are based on the "Big Five", also known as the Five Factor model of personality. The Five Factor model proposes that personality can be categorised into five broad factors: Extraversion, Agreeableness, Conscientiousness, Openness to Experience, and Neuroticism (for a comprehensive review see for example [12]).

Critically we are looking to explore the impact of expressive synthesis on user attitudes towards the voice and character in general. To test this we ran a controlled experiment where users were asked to rate a virtual character on a number of voice and character level dimensions. We varied the voice the character was given (either human recording or synthetic voice), keeping the graphical rendering method for the character constant in all conditions. The method used to obtain the synthetic speech in the study is described in the next section.

http://dx.doi.org/10.21437/Interspeech.2017-325

2. Speech Synthesis Method

2.1. Copy-Synthesis

Speech can be automatically generated from input text by using a parametric TTS system or can be synthesised by reconstructing the waveform from a recorded utterance, termed copy-synthesis. In either method, a parametric model of speech is commonly employed to represent the signal by parameters extracted from recorded speech. These parameters can include the pitch, duration and spectral parameters that depend on the configuration of the vocal tract articulators.

We use copy synthesis for the following reasons. The quality of synthetic speech is likely to be lower in TTS than in copy-synthesis because the first requires the additional part of predicting the speech parameters from the input text (using a machine learning algorithm), which may introduce perceptual speech distortions. Current TTS systems perform well for reading texts such as news, yet they have limitations in their ability to produce expressive speech in audiobooks or spontaneous dialogue contexts. For this reason, expressive TTS is usually limited to a few pre-defined basic emotions and voice styles. As we are interested in evaluating an expressive character to base the synthesis on we used copy-synthesis. This allowed us to maximise the expressivity in the synthetic speech, compared to parametric TTS. However, copy-synthesis can sound very close to human speech. In order to have a good trade-off between good reproduction of voice expressivity and a sufficient level of artificial sounding voice, we used a speech model that permits to have control on this dimension, as described in the next section.

Recent research developments in TTS such as WaveNet [13] and SampleRNN [14] show that very high-quality speech can be synthesised without the use of a speech vocoder. But these methods are still computationally expensive and in certain applications, particularly mobile applications with limited bandwidth internet connection, it may be advantageous to use a low memory footprint system such as the popular HTS [15] that sounds more artificial and depends on the quality of its vocoder. Moreover, the prosody of generated speech in state-of-the-art TTS systems is generally not of convincing quality compared with natural speech, especially in the expressive case. In this study we are not concerned in testing the limitations of the most advanced TTS systems in the context of expressive avatars. Instead, we want to pull apart the contributions of prosody modelling limitations in the evaluation of expressiveness and focus on the effects of the vocoder artefacts in the evaluation of the perception of the character. Obviously, the outcomes of our experiment cannot be generalised to other types of copy-synthesis methods and TTS systems, but they aim to provide initial insights in the study of the multi-dimensional and complex factors that affect perception of synthetic speech in this context.

2.2. Uniform Concatenative Excitation Model

As proposed in [16], we used the Uniform Concatenative Excitation Model (UCEM) as our speech model for producing the synthetic speech. It has the advantage of being robust to voicing classification errors. The model is a source-filter model that represents the source signal as the mix of a periodic signal with noise. Both the source and filter are estimated by performing pitch-synchronous Linear Prediction (LP) analysis to obtain the residual and the vocal tract filter. Then, for each pitch-mark (calculated using a glottal epoch detector), the periodic component is represented by a segment of the original residual around the pitch-mark (around 1.5 ms long). The remaining part of the residual during the pitch period is represented by a noise signal that is shaped by the amplitude envelope of the residual waveform and scaled in energy by using a measure of the power ratio between the two segments. The UCEM can be seen as an alternative model to the popular LPC vocoder [17], which uses a more refined model of the source component than the simple impulse/noise excitation used in LPC.

2.3. Speech Analysis and Synthesis

The parameters of the UCEM were estimated similarly as in [16]. First, the epochs were detected using the SEDREAMS algorithm [18] (detects peaks of the residual waveform in both voiced and unvoiced speech regions). Then, the start and end points of the segment representing the residual around the epoch were obtained by negative peak and a zero crossing detection, respectively. The amplitude envelope of the remainder part of the residual frame was represented by the local maxima in this segment and is parameterized by a non-linear polynomial fitting algorithm (order six). In addition to the LPC filter and polynomial coefficients the power ratio between the two segments was also calculated.

In the speech synthesis part, the excitation was generated by concatenating the the residual segment around the epoch with a white noise signal shaped in amplitude by the polynomial function and scaled in energy using the power ratio parameter. This method allows to control the degree of naturalness of the synthetic speech by adjusting the length of the residual segment in UCEM. This is an important advantage over other potential vocoders, such as the LPC vocoder. High-quality copy-synthesis can sound almost indistinguishable from the original speech so we decided to only use the periodic component of the source model in voiced sounds to obtain a more artificial voice. But we also adjusted a parameter of this model so that the length of the residual segment around the pitch-mark was sufficiently long so as to avoid excessive speech distortion. More precisely, we detected the start and end points of that component of the excitation on a segment centered around the epoch with duration of 1.5 ms instead of the 1.3 ms used in [16]. This value was selected based on perceptual judgments of the synthetic speech by one of the authors.

3. Experiment

3.1. Participants

31 participants (15 male, 16 female) took part in the study. Participants were staff and students of Trinity College Dublin and University College Dublin. They were recruited via email from a number of schools across both universities. Each participant was given a €5 voucher for participation.

3.2. Stimuli

In the experiment, participants watched eight video clips of around 10 seconds each that varied in the personality portrayed by the virtual character (shown in Figure 1). The clips showed the character expressing positive (extraversion, agreeableness, emotionally stability, openness) and opposite (introversion, non-agreeableness, neuroticism, non-openness) personality traits. We used short videos because we were interested in an immediate impression of the character.

The movement and voice of the character in each of the clips were based on the movements and voice of a male actor

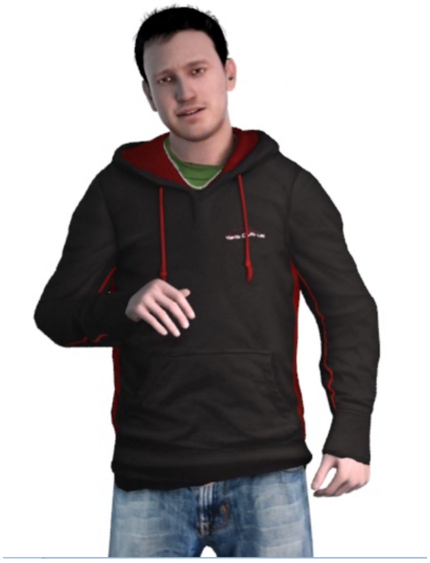

Figure 1: *Screenshot of the virtual character used in the study.*

asked to display such personality traits. To create the virtual character a 3D scan of the actor's face was taken using a structured light 3D scanner (ABW). The scan and photographs of the actor were then used by an artist to create the final virtual model with high quality diffuse, opacity and normal-map textures, and both facial and body rigs. The scene was rendered in 3ds Max 2015 with two area spotlights and ray-traced shadows. The character was similar to the HumanHQ1 condition seen in [19]. Further details can be found in [19].

The voice used by the character varied between subjects. Participants were either exposed to the virtual character using the natural recordings or a synthetic voice. Thus, in total 16 video clips were generated (8 per voice condition) for the experiment. The method used to obtain the synthetic speech was described in Section 2. The synthetic speech samples were normalized in energy to have similar level as the recorded speech, since differences in loudness can significantly affect speech perception.

3.3. Task

The experiment task focused around judging the male virtual character on a number of dimensions. Participants were asked to watch the eight clips. After each clip they were asked to rate aspects of the character in general and aspects specific to the character's voice. The questions they were asked after each video were:

- Character Appeal: How appealing did you find the character? (1=Not at all appealing; 7=Extremely appealing);
- Character Credibility: How credible did you find the character? (1=Not at all credible; 7=Extremely credible);
- Character Human-Likeness: How human-like did you find the character? (1=Not at all human like; 7=Extremely human-like);
- Voice Likeability: How much did you like the character's voice? (1=Not at all; 7=Very Much);
- Voice Consistency: How well did you think the voice matched the character's appearance? (1=Not at all well; 7=Extremely Well);
- Voice Expressiveness: How expressive did you find the character's voice? (1=Not at all expressive; 7=Extremely expressive);
- Voice Understanding: How easy did you find it to understand what the character was saying? (1=Not at all easy; 7=Extremely easy).

3.4. Procedure

Participants were welcomed into the lab, were given information about the study and were asked to complete consent forms and a short demographic questionnaire. They were then asked to watch the 8 short video clips. The sequence of the videos was randomised for each participant. After each clip they answered the questions indicated in the previous section. Finally, participants were debriefed as to the motivations of the study and thanked for taking part in the research. The study took approximately 10 minutes.

4. Results

Each measure was analysed using a 2 (Voice Conditions - between participants) $\times$ 8 (Video Conditions - within participants) Mixed Design ANOVA. We used Holm correction to control for Type 1 error in post-hoc comparisons [20]. This method is more powerful than traditional Bonferroni comparisons [21], especially when a large number of groups are being compared, as is the case in the Video independent variable.

We found that there were no statistically significant main effects ($p > .05$) of voice on the appeal, credibility and human-likeness dimensions measured. That is, people rated the virtual characters with each of the voices as similar in appeal, credibility and human-likeness. People also rated both voices as equally matching the appearance of the virtual character ($p > .05$). Yet we did obtain statistically significant main effects of voice for the other measures. First, people seemed to like the human voice more than the synthetic voice when judging the virtual character [$F(1, 29) = 14.90$, $p < .001$]. They also felt that the human recorded voice was more expressive than the synthetic voice used by the character [$F(1, 29) = 11.17$, $p = .002$]. Participants also stated that they found it easier to understand the human compared to the synthetic voice [$F(1, 29) = 30.12$, $p < .001$]. Figure 2 shows these results graphically. There were no interaction effects between videos and voice in any of the analyses conducted ($p > .05$) nor were there any statistically significant differences between the ratings for each video after controlling for Type I error in the post-hoc comparisons.

5. Discussion

This research contributes valuable insight into the role of voice in virtual character evaluation. We find that the voice conditions have a statistically significant effect on voice trait evaluations, in agreement with previous studies, e.g., [6]. However, we find that they have little impact on the character level judgments beyond those that are voice focused. It seems that, in this case, voice level features did not affect people's overall character level judgments.

Whilst we found natural recordings were judged as superior to synthetic speech, echoing previous work [6], we did not find a consistency effect for character evaluations seen in [6]. This

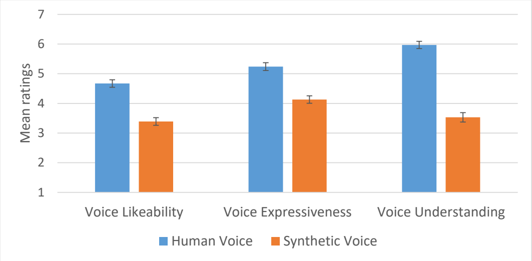

Figure 2: *Mean rating (with standard error bars) by voice type for Voice Likeability, Expressiveness and Voice Understanding.*

may be explained by our findings around voice consistency. Unlike in [6] our participants rated both voices as being consistent with the character's appearance, even though one voice was clearly more human-like than the other. They therefore saw both voices as being consistent with the rendering of the character, rather than the human voice being seen as less consistent to the virtual character. This may be because people are used to characters like ours having human recorded voices, especially in semi-realistic [23] computer-animated film and game characters. Both scenarios of a human recording and a synthetic voice may have therefore been seen as consistent in relation to the type of virtual character being judged.

The fact that people felt that the human voice recordings were more expressive, understandable and likeable than the synthetic voice used suggests that the distortion of the vocoded voice used had a significant impact on speech characteristics beyond naturalness. With the aim of improving expressive synthesis, it is critical that we determine the level of signal quality that is required for synthetic speech to be comparable to human speech. For instance, this could be done by evaluating how parameters of the UCEM model that affect levels of speech distortion influence user ratings. Indeed as expressive synthesis research develops it is important to compare different expressive copy-synthesis methods as well as extending the evaluation to current state-of-the-art TTS. From this we can then better understand how expressive synthesis brings benefits in terms of user perceptions and experiences.

6. Limitations and Future Work

Our work was focused on a specific domain, that of virtual characters, testing the role of voice in this context. Specifically the character was male and only one method of synthesis was studied. Future studies should look to test the impact of vocoded speech in other interaction contexts as well as testing different copy-synthesis methods and TTS systems. For instance, it may be that expressiveness affects users differently in situations where there is the potential for dialogue interaction between the character and the user, like in intelligent personal assistants and robot companion interaction, rather than in a one-way interaction tested here. Increasing expressiveness of synthesis may lead people to imbue more human-like traits onto such agents in conversational contexts [5, 4]. A fruitful avenue for further work could therefore look at the effects that expressive synthesis plays in user perceptions in such dialog interactions. Indeed the influence of voice and character gender as well as TTS methods on user ratings should also be explored.

Our research also focused on short-term exposure to such characters. The experiment focused on people's judgements of a virtual character in eight video clips of around ten seconds each. Further work should look to understand how speech synthesis affects judgments and the dynamic nature of judgments over more long-term interactions using a larger video dataset. Human judgments of such virtual characters are likely to be dynamic and change over time [3, 6]. These judgments may also be impacted not just by design cues such as voice type and graphical rendering, but also by the behaviour of the character and its interplay with speech and graphics. Similarly how speech expressiveness affects user behaviour should also be investigated. Our work focuses on subjective metrics, looking at how aspects of a character's voice affect user perceptions. Work by [6, 23] explore the impact of agent design (such as voice naturalness and humanness) on more objective aspects such as levels of agent interruption and a user's level of disclosure. The use of expressive synthesis may also have a strong impact on more objective measures such as user's behaviour in interaction. Similarly, the measure of human understanding could be extended with objective metrics, in future experiments. For instance, one could measure the word error rate in a transcription task [7] and speech comprehension (as quantified through question-answering or other task).

It is also important to consider the role that individual differences may play in people's evaluation of characters and synthesis more generally. Work on people's perceptions of synthetic voices found that people were more positive towards voices that matched their personality [24]. People's evaluation of characters can therefore be significantly affected by individual differences. Identifying these effects on synthesis evaluation more widely may be a fruitful avenue for future work.

7. Conclusions

Our work aimed to evaluate a synthetic voice compared to human voice, in the context of a talking virtual character. We looked to evaluate the voice on a number of characteristics such as likeability, expressiveness and understanding as well as how this voice affected character level evaluations such as the character's human-likeness, credibility and appeal. This paper contributes to the shift of speech synthesis evaluation into a wider number of variables when considering synthesis evaluation in audio-visual applications. We found that the people's judgments of the character's characteristics did not vary depending on the voice that the character used. In contrast, the voice did impact the avatar's communicative characteristics in that participants perceived the character as more understandable, expressive and liked their voice more when using the human than the synthetic voice.

8. Acknowledgements

This research is supported by Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie) at Trinity College Dublin, the UCD Seed Funding Scheme (Grant SF1191) and the Irish Research Council funded Cog-SIS project (Grant R17339).

9. References

[1] O. Türk and M. Schröder, “Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques,” in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 965–973, 2010.

[2] X. Li, Z. Wu, H. Meng, J. Jia, X. Lou and L. Cai, “Expressive Speech Driven Talking Avatar Synthesis with DBLSTM Using Limited Amount of Emotional Bimodal Data,” in *17th Annual Conference of the International Speech Communication Association INTERSPEECH*, San Francisco, USA, pp. 1477–1481, 2016.

[3] L. Gong, “How social is social responses to computers? The function of the degree of anthropomorphism in computer representations,” *Computers in Human Behavior*, vol. 24, no. 4, pp. 1494–1509, 2008.

[4] E. Luger and A. Sellen, “‘Like Having a Really Bad PA’: The Gulf Between User Expectation and Experience of Conversational Agents,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, pp. 5286–5297, 2016.

[5] B. R. Cowan et al., “What can I help you with?: Infrequent users’ experiences of Intelligent Personal Assistants,” *in Proceedings of Mobile HCI 2017*, Vienna, Austria, 2017.

[6] L. Gong and C. Nass, “When a Talking-Face Computer Agent is Half-Human and Half-Humanoid: Human Identity and Consistency Preference,” *Human Communication Research*, vol. 33, no. 2, pp. 163–193, Apr. 2007.

[7] S. King, “Measuring a decade of progress in Text-To-Speech,” *Loquens*, vol. 1, no. 1, 2014.

[8] S. King and V. Karaiskos, “The Blizzard Challenge 2016,” in *Proc. Blizzard Challenge Workshop*, 2016.

[9] W. J. Mitchell, K. A. Szerszen Sr, A. S. Lu, P. W. Schermerhorn, M. Scheutz and K. F. MacDorman, “A mismatch in the human realism of face and voice produces an uncanny valley,” *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[10] E. Mower, M. J. Mataric, and S. Narayanan, “Human Perception of Audio-Visual Synthetic Character Emotion Expression in the Presence of Ambiguous and Conflicting Information,” *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 843–855, 2009.

[11] K. Zibrek and R. McDonnell, “Does Render Style Affect Perception of Personality in Virtual Humans?,” *in Proceedings of the ACM Symposium on Applied Perception*, New York, NY, USA, pp. 111–115, 2014.

[12] G. Matthews, I. J. Deary and M. C. Whiteman, “Personality Traits,” Cambridge University Press, 2003.

[13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *arXiv* preprint 1609.03499, 2016.

[14] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, “SampleRNN: An unconditional end-to-end neural audio generation model,” in *Proc. ICLR*, 2017.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling,” *Proc. ICASSP*, pp. 229–231, 1999.

[16] J. P. Cabral, “Uniform Concatenative Excitation Model for Synthesising Speech without Voiced/Unvoiced Classification,” in *14th Annual Conference of the International Speech Communication Association INTERSPEECH*, Lyon, France, pp. 1082–1085, 2013.

[17] J. R. Deller, J. G. Proakis and J. H. Hansen, “Discrete Time Processing of Speech Signals,” *Macmillan*, New York, USA, 1993.

[18] T. Drugman and T. Dutoit, “Glottal Closure and Opening Instant Detection from Speech Signals,” in *10th Annual Conference of the International Speech Communication Association INTERSPEECH*, Brighton, U.K., pp. 2891–2894, 2009.

[19] R. McDonnell, M. Breidt, and H. H. Bülthoff, “Render Me Real?: Investigating the Effect of Render Style on the Perception of Animated Virtual Humans,” *ACM Trans. Graph.*, vol. 31, no. 4, pp. 91:1–91:11, 2012.

[20] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[21] M. Aickin and H. Gensler, “Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods,” *American Journal of Public Health*, vol. 86, no. 5, pp. 726–728, 1996.

[22] E. Lee, “The more humanlike, the better? How speech type and users’ cognitive style affect social responses to computers,” *Computers in Human Behavior*, vol. 26, no. 4, pp. 665–672, 2010.

[23] I. Gris, D. Novick, A. Camacho, D. A. Rivera, M. Gutierrez and A. Rayon, “Recorded Speech, Virtual Environments, and the Effectiveness of Embodied Conversational Agents,” in *Intelligent Virtual Agents*, pp. 182–185, 2014.

[24] C. Nass and K. M. Lee, “Does computer-generated speech manifest personality? an experimental test of similarity-attraction,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI ’00)*, New York, NY, USA, pp. 329–336, 2000.