Evolving emotional prosody

Cecilia Ovesdotter Alm[†], *Xavier Llorà*[‡]

[†]Department of Linguistics, [‡]NCSA
University of Illinois, Urbana-Champaign
ebbaalm@uiuc.edu xllora@illigal.ge.uiuc.edu

Abstract

Emotion is expressed by prosodic cues, and this study uses the *active interactive Genetic Algorithm* to search a wide space for SAD and ANGRY parameters of intensity, F0, and duration in perceptual resynthesis experiments with users. This method avoids large recorded databases and is flexible for exploring prosodic emotion parameters. Solutions from multiple runs are analyzed graphically and statistically. Average results indicate parameter evolution by emotion, and appear more distinct for SAD. Solutions are quite successfully classified by CART, with duration as main predictor.
Index Terms: emotions, prosody, interactive evolution, aiGA.

1. Introduction

Emotion is expressed by prosodic cues, but their interplay is an open question, which is complicated by a challenging search space. Common procedure for emotional speech research depends on recording and analyzing large databases. Instead, this work uses the *active interactive Genetic Algotihm* (aiGA) [1] to evolve emotional prosody in perceptual resynthesis experiments. Within this framework, fundamental parameters of emotional prosody become an optimization problem, approximated by searching perceptual space of listeners via interactive feedback. In contrast to unit or parameter estimation based on emotional speech databases, e.g. [2] [3], the method only requires NEUTRAL utterances as starting point, and user preferences guide the direction of the efficient aiGA. Thus, there is no need to record large emotion databases, and parameter findings are not limited to what is found in data; instead models evolve more freely, as permitted by imposed bounds. Results from an initial experiment on 1-word utterances with the goal to evolve ANGRY and SAD speech are analyzed, and indicate that aiGA evolves prosodic parameters by emotion. Solutions' emotion targets are also quite well predicted by a CART model.

2. Related work

Modifications in F0, intensity, and duration are facets of emotional prosody [4]. While ANGER is often characterized by increased speech rate, F0, and intensity, SADNESS is assumed marked by opposite behavior e.g. [5]. Other features have been suggested, but with less evidence, e.g. voice quality [6]. Synthesizing emotional speech has been attempted with various techniques [7]. EmoSpeak [8] allows manual manipulation of many parameters with an interesting dimensional interface, but parameters were fitted to a database and literature. An interesting study drew on a Spanish emotional speech corpus for Catalan emotional prosody [9].

Despite much previous work, emotional profiles remain unclear [10]. Fresh work may contribute to increased understanding of emotional prosody, and the suggested approach rephrases the research question as: on average, how is a particular emotion *best* rendered in synthetic speech? A step has been taken before toward structured search [11], but seemed to use a simple iGA, which ignores important considerations in interactive computation such as user fatigue and flexible targets [12] [13]. GAs [14] are iterative search procedures that resemble natural adaptation phenomena. Issues and applications in interactive evolutionary computation have been surveyed [12], as have recent advances in aiGA theory [13]. AiGA has been successful for speech, by interactively estimating cost functions for unit-selection TTS [1]; aiGA ensured high intra-run consistency in subjective evaluations and decreased user evaluations compared to a simple iGA, i.e. combating user fatigue.

Monosyllabic	sas, bem, face, tan
Bisyllabic	barlet, person, cherry, tantan
Trisyllabic	bubelos, strawberry, customer, tantantan

Table 1: Words used as resynthesis basis

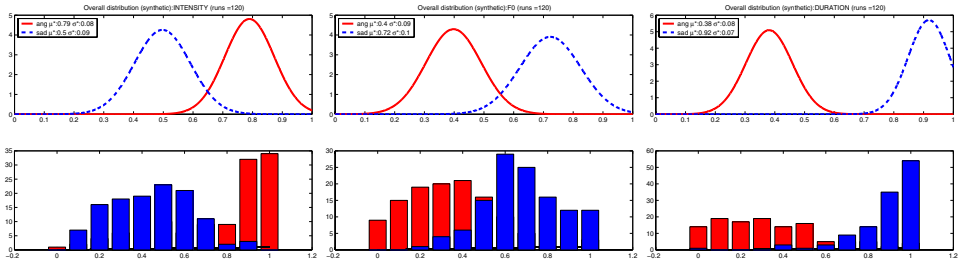
3. Experimental design

Interactive evaluation was used to evolve emotional prosody with aiGA developed by [13] [1]. In each run, a user's feedback guided the process to estimate performance and evolve a synthetic model beyond what was presented to the user (for details, cf. [1]). AiGA assumed variable independence and built a probabilistic model, in this case based on a population of normal distributions[1] with the $UMDA_c$ algorithm [16]. The output of each run r was an evolved synthetic normal model (μ_r, σ_r) for each prosodic variable.

In this experiment, users listened to and evaluated pairs of resynthesized utterances. The aiGA had parameter vectors or *individuals* with prosodic parameters for resynthesizing 1-word utterances. Each individual had 3 values for Int (intensity), $F0$ (mean F0), and Dur (total duration; i.e. word tempo) $\in [0, 1]$, encoded as proportions deviating from the original NEUTRAL word at 0.5, with truncation applied if the synthetic model evolved beyond $[0, 1]$. Each run had 3 iterations, which [1] found sufficient, and a user evaluated 22 pairs of sounds in a run. Individuals were initialized randomly with a different seed for each day of the experiment, except that one individual's values were set according to trends in the literature for each emotion (see sec. 2). The search space for each variable was delimited by upper and lower bounds, cf. Table 2, adjusted to the original voice to avoid unnatural speech.

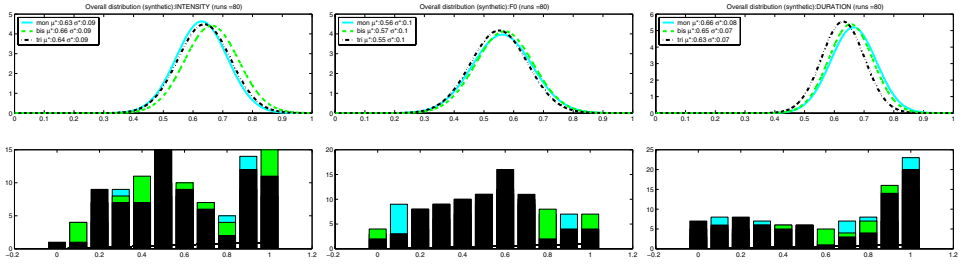
Conversion between actual values for resynthesis and their corresponding proportion in $[0, 1]$, as encoded for the aiGA by in-

[1]Listeners agree cross-culturally [15] above chance on ANGRY vs. SAD speech, which supports normality. The true distribution remains unknown.



(a) Intensity (b) F0 (c) Duration

Figure 1: Overall distributions (μ^*, σ^*) for intensity, F0, or duration show partly or fully separated curves by emotions SAD and ANGRY.



(a) Intensity (b) F0 (c) Duration

Figure 2: Overall distribution (μ^*, σ^*) for intensity, F0, or duration show overlapping curves for syllabic type.

Variable	Unit	Min	Max
Sound *Int*	dB	69	83
Mean *F0*	mel (Hz)	124 (139)	227 (282)
Total *Dur*	ratio	0.70 (shorter)	1.80 (longer)

Table 2: Upper and lower limits on word-level prosodic properties

dividuals' variables, was done with an approximation, where 0.5 represented the NEUTRAL original sound used as resynthesis basis, 0 corresponded to the minimum and 1 to the maximum allowed (cf. Table 2). Each user-evaluation was a tournament that involved listening to 2 resynthesized 1-word utterances, and selecting the one which the user felt best portrayed the target emotion, or indicating a draw. To further avoid user fatigue, the word for a tournament was chosen at random from a set of 4 NEUTRAL words, then resynthesized given 2 individuals' parameters, and the resulting pair of sounds presented to the user. The original words used as resynthesis basis, cf. Table 1, came from NEUTRAL declarative utterances recorded from a female US English speaker. Words were controlled for syllable length but not for segmental makeup, since emotional prosody should generalize beyond the micro-level.[2]

Resynthesis was done with two parallel Praat [17] implementations, and individuals were resynthesized on the fly in a stepwise process before each tournament, with the aiGA running in the background and regulating resynthesis parameters, user tournaments, and computation of performance. Except for variable input implementation, the model was constructed with future experiments on multiple-word utterances with local word-level encoding in mind. Thus, it involved separate resynthesis at the word level with a final concatenation-resynthesis component.

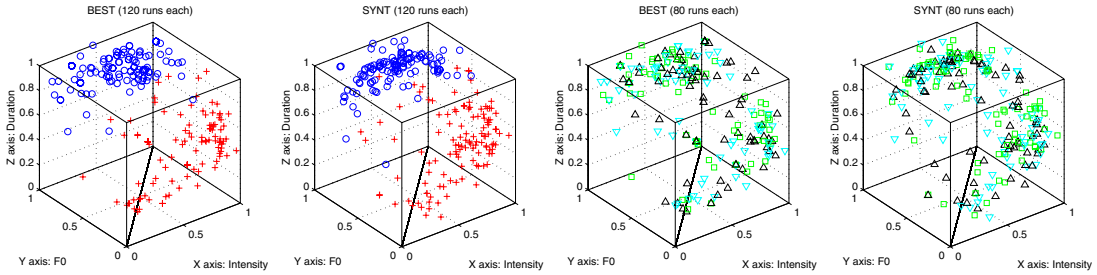
Interactive experiments involved 2 males and 2 females; all highly proficient in English, with either Swedish (3) or Spanish (1) as native language. Over 10 different days (within a 20-day period), they completed two blocks of 3 SAD tasks and 3 ANGRY task, with an intermediate short break, for each day, i.e. SAD and ANGRY target emotions in combination with either *monosyllabic*, or *bisyllabic*, or *trisyllabic* word types. The 10 day replicas were done to not overload users and keep them alert, and to reduce effects from random initialization or daily moods. Emotion perception is subjective, so averaged results across runs are of main interest. A web interface was used for user experiments. Post-feedback indicated contentment, but some felt 10 days was a bit long, or worried slightly about consistency or desensitization, or felt ANGRY was not as distinct as SAD. One person felt SAD had a "pleading" quality, and that some sounds reflected a nicer voice.

[2]Resynthesis is quite robust with some *F0* or *Int* inconsistencies (perhaps due to automatic tracking); deemed noninvasive by listening and compared to variability by user, voice, word. *Dur* could partly evolve with more stability. Also, due to client browser, a few runs were pruned and restarted.

4. Results and discussion

For each run r of the $4 * 10 * 6 = 240$ completed runs, the values (with proportion encoding) for *Int*, *F0* and *Dur* for its final best individual and final evolved synthetic model (i.e. evolved μ_r and σ_r) were extracted with a python script, with `matlab6.1` used for plotting and statistics. The data set of best individuals is henceforth called BI, and for evolved synthetic models ESM. The analysis intended to clarify that emotions' variables yielded distinct prosodic profiles, that aiGA was indeed evolving emotional prosody (i.e. not prosody by syllabic type), and what the averaged prosodic models were. The results representing the overall distribution of runs, based on ESM for the individual prosodic variables, are in Figs. 1 - 2, given proportion encoding $\in [0, 1]$ with truncation. The curves can be seen as representing the overall distribution (μ^*, σ^*) for the variables *Int, F0, Dur*, respectively, where $\mu^* = \frac{\sum_r \mu_r}{n}$, and $\sigma^* = \sqrt{\frac{\sum_r \sigma_r^2}{n}}$, where n is the number of runs r completed (e.g. for SAD runs $n = 120$, or for monosyllabic $n = 80$). The *pooled standard deviation* σ^* is an estimate of the larger population P (with unseen ANGRY/SAD cases). In contrast, the *sample standard deviation* s is larger, and this difference may be due to the number of runs being quite small. For completeness, histograms over the μ_r sample are also included.[3]

[3]Some columns may partially hide others.



(a) Emotion (SAD = ring, ANGRY = plus) (b) Syllabic (monos.=↓triangle, bis.=square, tris.=↑triangle)

Figure 3: Runs for BI (BEST) vs. ESM (SYNT) in 3D indicate trends for two clouds by emotion (a), but not by syllable type (b).

Tst#	Var.	Fact1	Fact2	#Rep.	Model	Sig. Fact1	Sig. Fact2	Sig. Interac.	`multcompare` diff. (sig. main Fact.)
1	Int	syl (3)	em (2)	40	BI		✓		SAD vs. ANG
1	Int	syl (3)	em (2)	40	ESM		✓		SAD vs. ANG
1	F0	syl (3)	em (2)	40	BI		✓		SAD vs. ANG
1	F0	syl (3)	em (2)	40	ESM		✓		SAD vs. ANG
1	Dur	syl (3)	em (2)	40	BI		✓		SAD vs. ANG
1	Dur	syl (3)	em (2)	40	ESM		✓		SAD vs. ANG
2	Int	user (4)	em (2)	30	BI		✓	✓	SAD vs. ANG
2	Int	user (4)	em (2)	30	ESM		✓	✓	SAD vs. ANG
2	F0	user (4)	em (2)	30	BI	✓	✓	✓	SAD vs. ANG; A vs. BCD
2	F0	user (4)	em (2)	30	ESM	✓	✓	✓	SAD vs. ANG; A vs. B
2	Dur	user (4)	em (2)	30	BI	✓	✓	✓	SAD vs. ANG; AC vs. BD
2	Dur	user (4)	em (2)	30	ESM	✓	✓	✓	SAD vs. ANG
3	Int	user (4)	syl-em (6)	10	BI		✓	✓	
3	Int	user (4)	syl-em (6)	10	ESM		✓		
3	F0	user (4)	syl-em (6)	10	BI	✓	✓	✓	
3	F0	user (4)	syl-em (6)	10	ESM	✓	✓	✓	(same as in test 2)
3	Dur	user (4)	syl-em (6)	10	BI	✓	✓	✓	
3	Dur	user (4)	syl-em (6)	10	ESM	✓	✓	✓	

Table 3: ANOVAs showed that *emotion* was always significant, but *syllabic type* was not. *User (persons A, B, C, D)* was significant for *F0*, *Dur*, with interactions. ✓ indicates significant p-values (Int = intensity, Dur = duration, syl = syllabic types, em = emotions, BI = final best individual, ESM = final evolved synthetic model, ANG = angry)

Fig. 1 shows that the overall distribution separates emotions, with some overlap for *Int* and *F0*, but not for *Dur*. As expected, the mean of *Dur* was shorter for ANGRY speech, and longer for SAD. For the mean of *Int*, the relative position of emotions to each other was as expected, but SAD was at the NEUTRAL middle. The mean of *F0* showed opposite behavior than the majority literature, with slightly decreased near NEUTRAL *F0* for ANGRY, but increased F0 for SAD. In contrast, syllabic types do *not* separate, cf. Fig. 2, and thus, do not seem to make a difference for average behavior.

When resynthesizing words with μ^* values, SAD appeared more distinct than ANGRY,[4] and ANGRY differed mildly from NEUTRAL, although certain words seemed angrier. Better SAD synthetic speech has been noted before [9]. The ANGRY emotion family may also be more diverse, and thus vary more.

Beyond isolated variables, Fig. 3(a-b) visualize runs in 3D as points in proportion encoding for 3 dimensions (*Int*, *F0*, and *Dur*) for BI and ESM (truncated μ_r values for ESM).[5] Despite outliers, and quite large estimated s for an emotion given its points and dimensions,[6] Fig. 3(a) indicates a trend of 2 clouds of points by emotion, which again contrasts with non-separation by syllabic type in 3(b). Albeit a run's ESM and BI points do not necessarily occur at same place, overall clouds seem similar for ESM and BI in 3(a).[7]

Next, for each prosodic variable, 2-way ANOVAs were done at 95% confidence level for data sets BI and ESM (with truncation), followed by a multiple comparison for significant main factors (using `matlab6.1`'s `anova2` and `multcompare`). Multiple comparison did not consider interactions and should be interpreted with caution. Results are in Table 3. The first test considered *syllable types and emotions*, and only the emotion factor showed significant difference. Interactions were not significant, and perceptual differences appeared due to emotion, and not to syllable type. The second test covered *users (persons A, B, C and D) and emotions.* Again, for all variables, emotion was a significant factor. For *F0* and *Dur* user was also significant, and interaction between factors was always significant. The third test regarded *users and emotion-syllable type task*. The emotion-syllable type task was a significant factor, and so were interactions (except for *Int* in ESM), as were users for, again, *F0* and *Dur*. Multiple comparisons showed that all tests grouped by emotion, and for the second and third tests, person A, a linguist expert, was involved when user was a significant factor. Feedback indicated A decided more analytically; novice users may be less "contaminated" by formal knowledge. However, user impact remains a point for further research since significant interactions were observed which are not yet well understood, and only 4 users were involved. Table 4 shows user behavior by emotion, prosodic variable (truncated μ_r for ESM), and data set, and indicates its complexity. Variation is quite noticeable, but *Dur* appears less varied for most subjects, at least for SAD.

CART (as implemented by M. Riley) was used on BI and ESM to see how far the binary distinction between SAD and ANGRY models obtained from runs could be learned, and to inspect what features supported prediction. Each example, labeled either SAD or ANGRY, had proportions for intensity, F0, and duration as features (non-truncated μ_r for ESM).[8] Mean precision, recall, and

[4] Resynthesis (incl. variation) by system with μ^* values: `http://www.linguistics.uiuc.edu/grads/ebbaalm/aiGApilot/aiGApilot_MU-STAR.zip`.

[5] Note as caution that 3D plots are merely descriptive, and may be visually misinforming due to dimensionality, scaling, or point overlap.

[6] For example, for BI $s_{sad} = 0.32, s_{ang} = 0.43$ when $s_{emotion_i} = \sqrt{s^2_{Int_{emotion_i}} + s^2_{F0_{emotion_i}} + s^2_{Dur_{emotion_i}}}$

[7] 16% of BI ANGRY equaled the individual set to literature values.

[8] Only 1 ESM fold had a decision node with value beyond $[0, 1]$ range.

BI-ANG	Person A	Person B	Person C	Person D	BI-SAD	Person A	Person B	Person C	Person D
Int	0.79 (0.21)	0.73 (0.27)	0.8 (0.24)	0.82 (0.22)	Int	0.41 (0.25)	0.58 (0.2)	0.48 (0.21)	0.46 (0.18)
F0	0.55 (0.27)	0.39 (0.25)	0.49 (0.25)	0.25 (0.19)	F0	0.78 (0.18)	0.64 (0.18)	0.63 (0.24)	0.84 (0.15)
Dur	0.24 (0.15)	0.4 (0.28)	0.23 (0.14)	0.47 (0.27)	Dur	0.92 (0.1)	0.97 (0.05)	0.94 (0.11)	0.89 (0.13)
ESM-ANG	Person A	Person B	Person C	Person D	ESM-Sad	Person A	Person B	Person C	Person D
Int	0.85 (0.2)	0.71 (0.3)	0.84 (0.21)	0.76 (0.26)	Int	0.44 (0.23)	0.56 (0.14)	0.47 (0.16)	0.51 (0.21)
F0	0.43 (0.16)	0.37 (0.23)	0.51 (0.25)	0.29 (0.18)	F0	0.77 (0.16)	0.65 (0.16)	0.66 (0.17)	0.81 (0.17)
Dur	0.35 (0.24)	0.39 (0.29)	0.26 (0.14)	0.53 (0.28)	Dur	0.9 (0.1)	0.98 (0.04)	0.94 (0.17)	0.85 (0.18)

Table 4: Users' means by emotion for BI and ESM (sample standard deviation in parenthesis; n=30 replicas).

Em-model	Mean prec.	Mean recall	Mean F	% non-unique exs.
ANG-ESM	0.90	0.88	0.88	0.05 (3 types)
ANG-BI	0.95	0.91	0.92	0.28 (7 types)
SAD-ESM	0.90	0.88	0.88	0.05 (3 types)
SAD-BI	0.92	0.94	0.93	0.26 (8 types)

Table 5: 10-fold cross validation means from CART classifying SAD and ANGRY evolved synthetic models (ESM) and best individuals (BI). Data set had 240 instances, i.e. 24 test examples in each fold. Mean results are well above the 50% baseline.

F-score based on 10-fold cross validation are in Table 5.[9] Interestingly, despite the sample variation, on average CART performed well above the 50% naïve baseline at distinguishing SAD and ANGRY instances. For ESM, 0.9 mean precision, 0.88 mean recall, and 0.88 mean F-score was obtained for both SAD and ANGRY predictions. For BI, performance even increased slightly, which may relate to BI having more repeated feature vectors, cf. col. 5 in Table 5. Inspection of decision trees showed that duration was mostly used as sole predictor. 5 ESM folds also used F0 for predictions, but intensity was not used. This may indicate a hierarchy of prosodic feature importance, and that some features may be subject to and show more vs. less variability; future work will clarify.

5. Conclusion

Given an initial study of 1-word utterances, the efficient aiGA was used to obtain average models of emotional prosody in interactive resynthesis experiments, with sadness appearing more distinct. At this point, microprosody and syllabic length appear less important, which supports word-level encoding, although some words seem better rendered than others with averaged solution, and user influence requires more work. Future experiments will include more users, longer utterances, and more emotions. Evaluating solutions for emotion recognition and naturalness could also be interesting. To conclude, aiGA has potential for evolving emotional prosody. Analysis indicated that average intensity, F0 and duration behaved differently for the 2 emotions, and F0 showed an interesting opposite behavior than expected. Moreover, 3D plotting indicated trends by emotion, and CART models showed that emotion solutions across runs were predictable to quite high degree, with duration appearing most indicative for prediction.

6. Acknowledgements

Many thanks to C. Shih, R. Proaño, D. Goldberg, and R. Sproat for valuable comments. The work was funded by NSF ITR-#0205731, the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under FA9550-06-1-0096, and the NSF under IIS-02-09199. Opinions, findings, conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the funding agencies.

7. References

[1] F. Alías, X. Llorà, L. Formiga, K. Sastry, and D. Goldberg, "Efficient interactive weight tuning for TTS synthesis: reducing user fatigue by improving user consistency," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, 2006, vol. 1, pp. 865–868.

[2] Richmond K. Hofer, G. and R. Clark, "Informed blending of databases for emotional speech synthesis," in *Interspeech*, 2005, pp. 501–504.

[3] C. Hsia, C. Wu, and T. Liu, "Duration-embedded bi-HMM for expressive voice conversion," in *Interspeech*, 2005, pp. 1921–1924.

[4] K. Scherer, "Vocal communication of emotion: a review of research paradigms," *Speech Comm.*, vol. 40, no. 1-2, pp. 227–256, 2003.

[5] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," *J. Ac. Soc. of Am.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[6] A. Ni Chasaide and C. Gobl, "Voice quality and the synthesis of affect," in *COST 258*, 2001, pp. 252–263.

[7] M. Schröder, "Emotional speech synthesis: A review," in *Eurospeech*, 2001, pp. 561–564.

[8] M. Schröder, "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions," in *ADS*, 2004, pp. 209–220.

[9] I. Iriondo, F. Alías, J. Melenchón, and A. Llorca, "Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis," in *ADS*, 2004, pp. 197–208.

[10] M. Tatham and K. Morton, *Developments in Speech Synthesis*, Wiley, 2005.

[11] Y. Sato, "Voice quality conversion using interactive evolution of prosodic control," *App. Soft Comp.*, vol. 5, pp. 181–192, 2005.

[12] H. Takagi, "Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation," in *IEEE*, 2001, vol. 89, pp. 1275–1296.

[13] X. Llorà, K. Sastry, D. Goldberg, A. Gupta, and L. Lakshmi, "Combating user fatigue in iGAs: partial ordering, support vector machines, and synthetic fitness," Tech. Rep. IlliGAL No 2005009, UIUC, 2005.

[14] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading: Addison W., 1989.

[15] Å. Abelin and J. Allwood, "Cross linguistic interpretation of expression of emotions," in *8th Int. Symposium on Social Communication*, 2003.

[16] C. Gonzáles, J. Lozano, and P. Larrañga, "Mathematical modelling of UMDAc algorithm with tournament selection: Behaviour on linear and quadratic functions," *Int. J. of Approx. Reasoning*, vol. 31, pp. 313–340, 2002.

[17] P. Boersma and D. Weenink, "`Praat`: doing phonetics by computer (version 4)," http://www.praat.org, Summer 2005.

[9] With $\frac{9}{10}$ train vs. $\frac{1}{10}$ test, with a different tenth for test in each fold, using the default parameters (except minor tree cutoff to avoid an overfit).