Social and Linguistic Speech Prosody

Proceedings of the 7th international conference on Speech Prosody

SpeechProsody 7

(Trinity College Dublin) May 20-23, 2014

Fondúireacht Eolaíochta Éireann
Science Foundation Ireland

ISSN: 2333-2042

“Young” and “Old” Voice: the prosodic auto-transplantation technique for speaker’s age recognition

Massimo Pettorino, Elisa Pellegrino, Marta Maffia

Department of Literary, Linguistic and Comparative Studies,
University of Naples “L’Orientale”, Italy
{mpettorino, epellegrino, mmaffia}@unior.it

Abstract

The present study is intended to figure out the extent to which prosody and intonation affect listeners’ ability to estimate the speaker’s age. The performance of a 40-year old anchorman and another by the same speaker at the age of 80 were spectro-acoustically analyzed in order to identify the prosodic features of a “young” and an “old” voice. The results of the analysis have shown significant differences between the two voices on a suprasegmental level. To test the effects of these differences on a perceptual level, through the prosodic transplantation technique, the F0 values and the durations of segments and silences were transferred from the “young” to the “old” voice and viceversa. Two age recognition tests, based on original and transplanted voices, were administered to Italian listeners. The results of perceptual tests have confirmed the strict relationship between some rhythmic and prosodic features and the speaker’s age and have demonstrated the effectiveness of the transplantation technique. With advancing age, articulation rate and speech rate slow down, voice register rises and tonal range widens. Moreover, the “old” voice is also characterized by a higher percentage of vocalic portion which determines a shift of the Italian rhythm towards the isomoraic pattern.

Index Terms: Prosodic correlates of speaker's age, Speaker's age recognition, Prosodic Transplantation Technique.

1. Introduction

The relationship between the speaker’s age and his/her own voice has been investigated in several experimental studies. Acoustic [1], [2], [3] and perceptual researches [4], [5], [6] have underlined that the voice changes with age, at both segmental and supra-segmental levels, because of the numerous anatomical and physiological modifications of the respiratory mechanism and of the phonatory apparatus. For example, lung tissue loses its elasticity, the thorax tends to stiffen, muscles weaken and vital capacity decreases [7]. In the elderly, a process of calcification of laryngeal cartilages also occurs and the vocal folds become thinner, stiffer and less elastic [8]. Previous studies on the effects of aging on acoustic parameters of voice have demonstrated that older voices are likely to undergo progressive tonal lowering [9], lowering of speech rate [10], [1], increasing of jitter and shimmer [8], [11], [12], lowering of formant frequencies [13], lengthening of vowels and stop consonants [14], increasing of standard deviation of F0 [13], [15], [16].

Nevertheless, it is important to underline that these data result from studies that analyze different kinds of corpora (spontaneous or read speech), use different techniques of speech data collection and involve different languages [2]. The considerable methodological heterogeneity across these kinds of studies certainly does not guarantee that the changes undergone by older voices are dependent exclusively on the speaker’s age. Besides the chronological age, other relevant variables, such as the idiosyncratic characteristics of a speaker's voice, the contextual situation, the kind of speech and the subject matter could influence the oral performance of the speakers and affect data comparability.

In order to control all these variables, and thus, to be able to exclusively assess the effect of aging on the acoustic parameters of voice, it would be very effective to record the same speakers uttering the same words in the same communicative situation, but at different ages.

2. The study

The objective of this study is to verify the role played by prosody and intonation in the listener’s ability to evaluate the speaker’s age. To achieve this, a particular corpus of Italian speech was collected. The read speech of a 40-year old anchorman, Piero Angela, was extracted from a 1968 TV news broadcast and orthographically transcribed. In 2007, the 79-year old Piero Angela, who still worked as a RAI journalist, read the same 1968 script again, acting as if he were hosting a real TV news broadcast. The recording was taken at RAI TV studios in Rome, in order to maintain the same communicative situation.

Preliminary spectro-acoustic analyses conducted on the 1968 and 2007 corpora showed differences between the two voices, both on the segmental and suprasegmental levels [17]. The “old” speech was clearly more isochronous, exhibited wider tonal range and presented longer and more frequent silent pauses than the “young” voice.

In order to investigate further the role played by the specific acoustic parameters of the “young” and “old” voices, three utterances drawn from the 1968 TV news broadcast and the corresponding utterances of the 2007 corpus were manipulated through the prosodic transplantation technique. This procedure is based on the PSOLA (Pitch-Synchronous Overlap and Add) algorithm and it is implemented in Praat [18], [19], [20]. The six utterances were segmented and annotated into four tiers:

- phones
- syllables
- intervals of consecutive consonants and vowels
- intervals between two consecutive vowel onset points (VtoV).

In order to apply the transplantation procedure, the “phones” tiers were duplicated and modified, so that each segment of the 1968 utterance had a corresponding segment in the 2007 utterance. Since the transplantation procedure requires that the TextGrids of donor’s and receiver’s voices contain the same number of elements [21], a micro segment was inserted to avoid mismatch between the 1968 and 2007 utterances. Thanks to the transplantation procedure, the rhythmic and prosodic features of the donor’s voice were transferred to the receivers' voice. As a result, in the present study, the utterances produced in 1968 by the 40-year old Piero Angela were made sound "older" by transferring the pitch contour on them, the durations of phones and silences of the corresponding 2007 utterances. By contrast, the latter were made to sound "younger" since they acquired the prosodic features of the 1968 utterances.

The resulting corpus was, thus, made up of three original utterances produced in 1968, three original utterances produced in 2007, three 1968 utterances with 2007 prosody and three 2007 utterances with 1968 prosody.

2.1. Perception test

In order to assess on a perceptual level the effect of the acoustic differences between the "old" and "young" voices, 70 university Italian students, ranging in age from 23 to 26, were administered a perception test. Following the experimental protocol used in a previous research [6], the test was divided into two main phases. In the first phase, participants listened to 16 pairs of utterances. Each pair was composed of two voices reading the same news. For each pair, listeners were asked to rate if the youngest speaker was the "1st voice" or the "2nd voice", or if the two speakers were of the "same age". The test was designed with the following voice combinations:

- 2007 voice - 2007 voice (Old –Old; henceforth O-O);
- 1968 voice -1968 voice (Young-Young; henceforth Y-Y);
- 2007 voice – 1968 voice (O-Y);
- 1968 voice - 2007 voice with the transplanted prosody of 1968 voice (Y-tY);
- 2007 voice - 2007 voice with the transplanted prosody of 1968 voice (O-tY);
- 1968 voice - 1968 voice with the transplanted prosody of 2007 voice (Y-tO);
- 2007 voice - 1968 voice with the transplanted prosody of 2007 voice (O-tO);
- 2007 voice with the prosody of 1968 voice – 1968 voice with the prosody of 2007 voice (tY-tO).

The test material also included a control pair (cO-cY), consisting in a 85 year-old voice and a 27 year-old voice reciting the same utterance: "le foglie diventano gialle, l'albero muore" (Engl. "when the leaves turn yellow, the tree dies"). In the second phase of the perception test, the same subjects were asked to listen to 14 single utterances and to indicate the speaker's age, choosing between seven age bands: 26-35, 36-45, 46-55, 56-65, 66-75, 76-85, 86-95.

2.2. Results of perception test

2.2.1. First test

Overall results of the perception test show listeners' ability to accurately recognize the speaker's age from speech sample alone. As a matter of fact, when exposed to the pairs based on the same voice O-O, Y-Y, these pairs are properly rated as being of "the same age", respectively in 100% and 96% of cases. When the pairs were composed of the 2007 voice and the 1968 voice (O-Y), 84% of listeners rate the second voice as the youngest.

The results obtained by the pairs composed by original and transplanted voices confirm the role of prosody and intonation in the perception of speaker's age. At the same time, the listeners' judgments highlight the effectiveness of the prosodic transplantation technique in making voices sound "older" or "younger". As for the aging effect, it was tested both with respect to the 2007 "old" voice and to the 1968 "young" voice (figs 1 and 2). As it is shown in figure 1, when the aged voice (tO) was paired with the original old voice (O), on average, the 60% of listeners judged the two items as being of the "same age".

Comparing the rates given to the O-tO pair with those assigned to the O-Y pair, the percentage of the answer "2nd voice" decreases from 84.2% to 30.8%, while the answer "same age" increases by about 50%, from 13 to 60.3%.

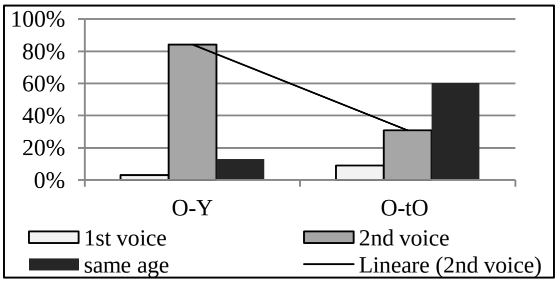

Figure 1: Aging effect with respect to the 2007 old voice. The rates given to pairs O-Y and O-tO are significantly different (p<0.01).

The data in figure 2 show the results obtained from the comparison between the young and the aged voice (Y-tO). 60% of listeners rate the "1st voice" as the youngest. The aging effect is particularly evident if one compares the answer "same age" obtained with the Y-Y pair and that of the Y-tO pair. In the former case, the third option is chosen by 96.8% of listeners, while in the latter, this percentage decreases to 38%.

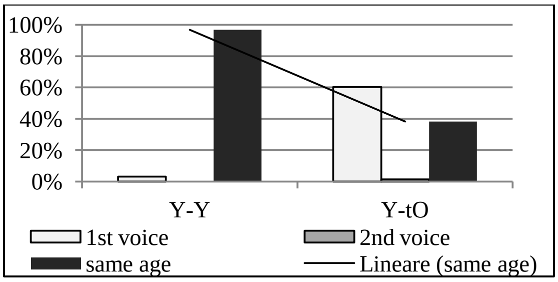

Figure 2: *Transplantation effect with respect to the 1968 young voice. The rates given to pairs Y-Y and Y-tO are significantly different (p<0.01).*

As regards the effect of making the voice sound "younger" (fig. 3), it is possible to claim that it produces changes in the "old" voice but its influence on listeners is not as effective as the aging effect. In the pair O-tY, the transplanted voice is recognized as the youngest by the 31.8% of listeners.

The data in figure 4 show that the rates given to the two pairs Y-O and Y-tY do not undergo a significant variation (p>0.05). However, in the Y-tY pair the percentage of "1st voice" decreases from 84.2 to 67. By contrast, the percentage of "same age" increases from 13 to 20.5 and in 11.5% of cases the synthesized young voice is even recognized as the youngest. In the pair composed of both transplanted voices, tY-tO, 56% of listeners judge it as if it was composed of voices of the same age.

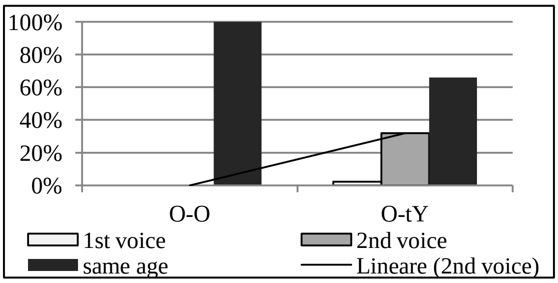

Figure 3: *Transplantation effect in respect to the 2007 old voice. The rates given to pairs O-O and O-tY are significantly different (p<0.01)*

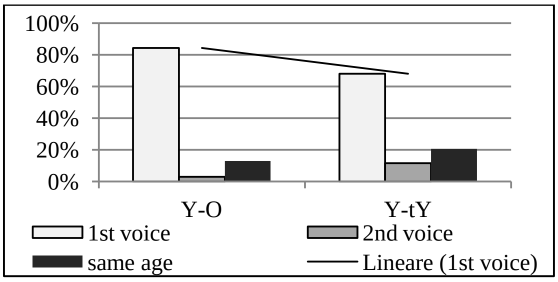

Figure 4: *Transplantation effect in respect to the 1968 young voice.*

2.2.2. Second test

In the second perception test, listeners were asked to estimate the speaker's age, choosing from 7 age bands. The data in figure 5 show that the young control voice (actual age 27) is recognized as belonging to the first age band (26-35) by 98% of listeners; while the control old voice (actual age 85) is judged by the 60% as ranging in the band 76-85. On a perceptual level, the 1968 voice varies from the 2nd to the 4th band, with the highest percentage of rates given to the 3rd band (46-55). On the contrary, the rates given to the 2007 voice are distributed on higher age-bands (4-6), with 44% of listeners choosing the 5th band.

As for the synthesized voices, they occupy an intermediate position between the original young and old voices. The age bands chosen by the highest number of listeners were the 4th and the 5th. These results indicate that the prosodic transplantation technique had a strong effect on the recognition of speaker's age in both directions: the aged voice is perceived as older than the original 1968 voice, as well as the synthesized young voice being rated as younger than the original old voice. From the data relative to the age bands, the weighted average of the perceived age was calculated, taking into account the recognition percentage scored by each band. Figure 6 plots the perceived and actual ages of the original, control and transplanted voices. The comparison between the perceived and actual ages of the speakers suggests different considerations. First of all, the variance between the two original voices decreases from the actual 40 years to perceived 20 years. The manipulated voices are perceived as produced by 66 and 65-year old speakers respectively.

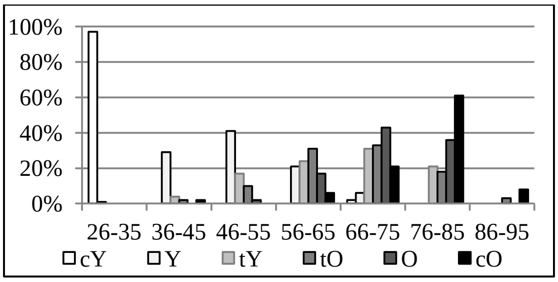

Figure 5: *Perceived age of control, original and transplanted voices.*

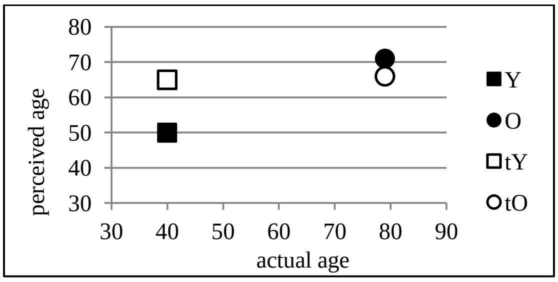

Figure 6: *Perceived and actual age of the original and transplanted voices.*

2.3. Spectro-acoustic analysis

The results of perception have shown that the 70 subjects were able to accurately judge speaker's age relying solely on single voice excerpts. This datum leads us to assume that the perception of the speaker's age is greatly influenced by the prosodic and intonational features of the utterance. To test this hypothesis, the whole corpus was spectro-acoustically analyzed. On the basis of the speech segmentation in phones, syllables, vocalic and consonantal intervals, VtoV intervals, the following measurements were taken: duration of phones, syllables, silent and filled pauses, length of vocalic and consonantal intervals, duration of VtoV intervals. Additionally, for every utterance the minimum, the maximum and the mean F0 values were measured.

On the basis of these measurements, the following rhythmic and prosodic parameters were calculated:

- articulation rate (AR) (syll/s.), i.e. the ratio between the number of syllables really uttered and phonation time;
- speech rate (SR) syll/s), i.e. the ratio between the number of syllables really uttered and total time of the utterance, including silent and non silent pauses;
- fluency (F) or frequency of silences in the utterance;
- tonal range (Hz);
- F0 register (Hz);
- speech time composition in terms of syllable, silence and disfluence percentage;
- utterance composition in terms of vocalic and consonantal percentage (%V and %C);
- mean duration of VtoV intervals (s).

In order to figure out which of the above-mentioned prosodic parameters could have influenced the listener's discrimination between the “young” and the “old” voice more profoundly, for every prosodic parameter the average values obtained in the 1968 corpus and 2007 corpus were calculated.

Table 1 shows the data regarding AR and SR. Unsurprisingly, the 2007 voice exhibits a decrease both in terms of AR and SR, due to the slower mobility of the articulators. However in both the 1968 and 2007 voices, the SR is 0.8 syll/s slower than AR, and this is imputable to the same portion of silent pauses (13%) occurring in the two corpora. It is worth underlining that, despite the duration of the silent portion being stable between the two voices, the frequency of silences increases considerably with advancing age: in 1968 silences are more rare, on average 1 out of 14 syllables, while in 2007 they are more frequent, 1 out of 7.6 syllables.

Table 1. *AR and SR.*

	AR	SR
1968	6.3	5.5
2007	5.4	4.6

As for pitch contour, figure 7 shows that with advancing age the voice register and tonal range become higher. In the 2007 voice the average pitch reaches 146,8 Hz, while it is lower in the 1968 voice (68 Hz). The tonal range widens by about 17% from 1968 to 2007.

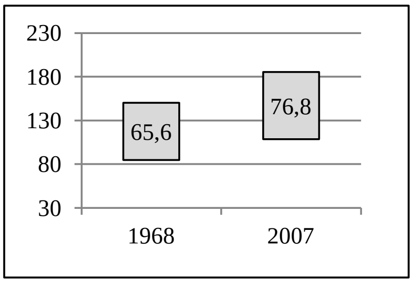

Figure 7: *Tonal range and Register.*

The other parameter under investigation was the vocalic and consonantal percentage in the utterance. Likewise the pitch movements, as well as the vowel percentage in the utterance increases with advancing age. In the corpus produced when Piero Angela was 40 years old, the vocalic portion amounts to 46%, while in the speech of the 80-year old speaker it reaches 51%. This datum enables us to clarify the components – vowels rather than consonants - on which the AR slowing down mostly depends. This phenomenon can be explained by the greater articulatory stability that characterizes vowels rather than consonants. The AR slowing down is, therefore, mainly due to the maintenance of static positions rather than to a variation of articulatory dynamics.

The variations of vocalic portion deserve more attention since this parameter plays a significant role in the rhythmic classification of languages. Indeed, according to a number of studies conducted on different languages, the traditional division in three rhythmic groups (syllable, stress, mora-timed languages) is related to vowel percentage and to another parameter, that according to [22] corresponds to ΔC (standard deviation of consonantal portions) and according to [23] to VtoV. In order to process the data of this corpus in the research framework of rhythmic organization of languages, the average %V and VtoV values obtained in Piero Angela's corpus were plotted with the ones obtained by [23]. Figure 8 shows the overall data relative to the different corpora. As it is clearly shown in the graph, the utterances produced by the 40-year old speaker lie in correspondence to the Italian TV news broadcast. On the contrary the mean VtoV and %V values of the utterances produced by the 80-year old speaker move to the right side of the graph, in the area occupied by Japanese, a mora-timed language. Thus, with advancing age, Italian speech seems to assume values more similar to mora-timed languages.

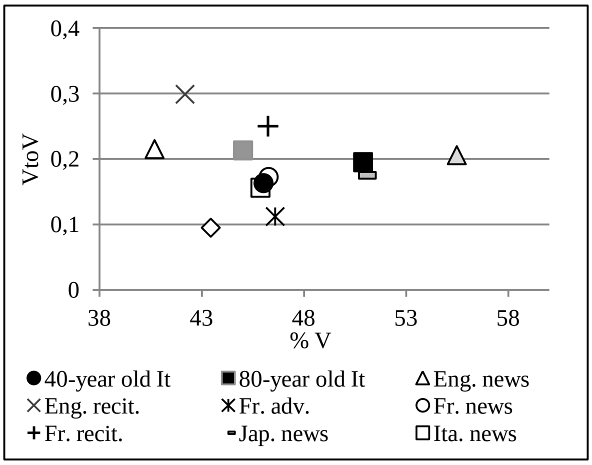

Figure 8: *%V and VtoV of different languages.*

This datum does not contrast the results of the above mentioned studies on the rhythmic features of languages, since neither the speakers involved in [22] nor the ones considered in [23] were 80 years old. Preliminary data on the speech produced by Japanese speakers of different age seem to confirm the trend which emerged for Italian: %V increases with advancing age (from 50.7 to 55.4%).

3. Conclusions

The spectro-acoustic analysis carried out on the corpus of read TV news broadcasts, produced by the same speaker at the age of 40 and 80 has demonstrated the existence of a relation between some rhythmic and prosodic features and the speaker's age. The old voice was characterized by a slowing down of AR and SR, by the rising of voice register and the widening of tonal range. The old voice also presents a higher percentage of vocalic portion (%V), and this increase makes Italian rhythm become more similar to the pattern of isomoraic languages. Thanks to the prosodic transplantation technique, the effect of these differences on a perceptual level was assessed. The judgments given to the transplanted voices suggest different considerations. Firstly, listeners are able to recognize the speaker's age relying solely on specific prosodic features. Secondly, on a perceptual level, the aging effect is more effective than that of making voices sound younger. In addition, the transplantation technique is therefore an effective procedure to the purpose of manipulating a speaker's age.

The data regarding the %V increase with advancing age suggesting the opportunity to extend the research on other languages belonging to other rhythmic groups.

4. References

[1] Hollien, H., Shipp, T., "Speaking fundamental frequency and chronological age in males", *Journal of Speech and Hearing Research,* 15, 155-159, 1972.

[2] Russell, A., Penny, L. and Pemberton, C., "Speaking fundamental frequency changes over time in women: A longitudinal study", *Journal of Speech and Hearing Research,* 38, 101-109, 1995.

[3] Schötz, S., "Acoustic Analysis of Adult Speaker Age", in C. Müller [Ed.], Speaker Classification I, Lecture Notes in Computer Science, 88–107, Springer, 2007.

[4] Horri, Y., Ryan, W. J., "Fundamental frequency characteristics and perceived age of adult male speakers", Folia Phoniatrica, 33, 227-233, 1981.

[5] Schötz, S., "Perception, analysis and synthesis of speaker age", *Travaux de l'Institut de Lund*, 47, 1-186, 2006.

[6] Pettorino M., Giannini A., "The speaker's age: a perceptual study", *Proceedings of the 17th ICPhS,* Hong Kong, 1582-1585, 2011

[7] Awan, S.N., "The aging female voice: acoustic and respiratory data", *Clinical Linguistics & Phonetics,* 20/2-3, 171-180, 2006.

[8] Linville, S.E., "The aging voice", *The American Speech-Language-Hearing Association (ASHA) Leader,* 12-21, 2004.

[9] Lindblad, P., *Rösten*, Lund, Studentlitteratur, 1992.

[10] Amerman, J.D., Parnell, M.M., "Speech timing strategies in elderly adults", *Journal of Phonetics*, 20, 65-76, 1992.

[11] Ramig, L.A., Ringel, R.L., "Effects of physiological aging on selected acoustic characteristics of voice", *Journal of Speech and Hearing Research*, 26, 22-30, 1983.

[12] Dehqan, A., Scherer, R. C., Dashti, G., Ansari-Moghaddam, A., & Fanaie, S., "The effects of aging on acoustic parameters of voice", *Folia Phoniatrica et Logopaedica*, 64, 265-270, 2012.

[13] Linville, S.E., "Acoustic-perceptual studies of aging voice in women", *Journal of Voice*, 1, 44-48, 1987.

[14] Ptacek, P.H., Sander, E.K., "Age recognition from voice", *Journal of Speech and Hearing Research*, 9, 273-277, 1966.

[15] Jacques, R., Rastatter, M., "Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners", *Folia Phoniatrica*, 42, 118–124, 1990.

[16] Traunmüller, H., van Bezooijen, R., "The auditory perception of children's age and sex", *ICSLP,* 1171-1174, 1994.

[17] Giannini, A., Pettorino, M., "L'età della voce", *La Fonetica Sperimentale: Metodo e Applicazioni, Atti del IV Convegno Nazionale AISV 4*, 165-178, 2009.

[18] Charpentier, F. and Moulines, E., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication* 9, 453-467, 1990.

[19] Boersma, P., "Praat, a system for doing phonetics by computer", *Glot International* 5:9/10, 341-345, 2001.

[20] Yoon, K. "Imposing Native Speakers' Prosody on Non-native Speakers' Utterances: The Technique of Cloning Prosody", *Journal of the Modern British & American Language & Literature* 25(4): 197-215, 2007.

[21] Pettorino, M. and Vitale, M. "Transplanting native prosody into second language speech", in M. G. Busà and A. Stella [Eds], *Methodological Perspectives on Second Language Prosody*. Papers from ML2P 2012, 11-16, Padova: CLEUP, 2012

[22] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", *Cognition* 73, 265-292, 1999.

[23] Pettorino, M., Maffia ,M., Pellegrino, E., Vitale, M. and De Meo, A., "VtoV: a perceptual cue for rhythm identification" in Mertens, P. & A.C. Simon [Eds], *Proceedings of the Prosody-Discourse Interface Conference 2013* (IDP-2013). Leuven, September 11-13, 2013, 101-106, 2013.